Modern Dynamic Programming Approaches to Sequential Decision Making

Seungki Min

Submitted in partial fulfillment of the
requirements for the degree of
Doctor of Philosophy
under the Executive Committee
of the Graduate School of Arts and Sciences

COLUMBIA UNIVERSITY

2021

Abstract

Modern Dynamic Programming Approaches to Sequential Decision Making

Seungki Min

Dynamic programming (DP) has long been an essential framework for solving sequential decision-making problems. However, when the state space is intractably large or the objective contains a risk term, the conventional DP framework often fails to work. In this dissertation, we investigate such issues, particularly those arising in the context of multi-armed bandit problems and risk-sensitive optimal execution problems, and discuss the use of modern DP techniques to overcome these challenges such as information relaxation, policy gradient, and state augmentation. We develop frameworks formalize and improve existing heuristic algorithms (e.g., Thompson sampling, aggressive-in-the-money trading), while shedding new light on the adopted DP techniques.

Table of Contents

Acknowledgements

I would like to express my deep gratitude to the following people, without whom I would not have been able to complete my doctoral study. I thank my supervisors, Prof. Ciamac Moallemi and Prof. Costis Maglaras, for providing invaluable guidance in all aspects, the faculty members in the Decision, Risk, and Operations division, especially Prof. Daniel Russo, Prof. Yash Kanoria, Prof. Santiago Balseiro, and Prof. Mark Broadie, for sharing inspiring ideas, and my dear colleagues, Pengyu Qien, Muye Wang, and Alex Yu, for their friendship and encouragement. Finally, my special thanks go to my wife Jiyun Kim and our newborn daughter Chaewon Min.

Chapter 1: Introduction and Background

Dynamic programming (DP) has long been an essential framework for solving sequential decision-making problems in a wide variety of domains. In a DP framework, the stochastic environment that the decision maker (DM) encounters is described by a Markov decision process (MDP), and the optimal policy that maximizes the expected cumulative reward (or minimizes the expected cumulative cost) can be obtained by solving the Bellman equation analytically or numerically.

Despite its wide applicability, the DP approach often presents challenges in real-world applications. When the state space is continuous or intractably large, for example, the optimal policy cannot be implemented unless the Bellman equation admits an analytic solution. When the DM is not risk-neutral (e.g., the DM's objective is not just the expected reward but also contains some risk term), the Bellman's optimality principle is no longer valid. In what follows, we illustrate such challenges that arise in multi-armed bandit problems and risk-sensitive optimal execution problems, and discuss how to overcome these issues using modern DP techniques.

1.1 Bayesian multi-armed bandit problem and Thompson sampling algorithm

The multi-armed bandit (MAB) problem concerns a situation where the DM is given a set of arms with unknown reward distributions and decides which arm to select at each time so as to maximize the cumulative reward. This problem specifically highlights the issue that the DM has to find a balance between exploitation (i.e., selecting the currently known best arm) to maximize the immediate reward and exploration (i.e., selecting an arm that has not been tested enough) to maximize the informational gain. As the simplest instance of a reinforcement learning problem, the MAB problem has received enormous attention over the past decades.

In the earliest work in the bandit literature, the MAB problem was considered in a Bayesian setting and formulated as an MDP problem, in which the DM's belief on the unknown model parameters is interpreted as a state that evolves according to the Bayes' rule whenever the DM observes a reward realization. Under this MDP formulation, the optimal policy exists as a solution to the associated Bellman equation; e.g., the seminal work of [1] characterizes such an optimal policy in a discounted infinite horizon setting. However, the optimal policy is not feasible to implement in most cases, since the belief state space is intractably large (its size scales exponentially in the number of arms) and the Bellman equation cannot be solved explicitly.

We particularly focus on Thompson sampling (TS), which we understand as an approximate DP solution that effectively mitigates the curse of dimensionality discussed above. TS utilizes the idea of "posterior sampling"; i.e., at each decision epoch, it draws a random sample of the model parameters from the posterior distribution and selects the arm that is best given the sampled model parameters, i.e., it makes a decision as if the sampled model parameters are the ground truth. Due to its intuitive mechanism and computational efficiency, it has been enjoying tremendous success in practice and is being adopted and implemented by Google, Microsoft, Facebook, and many other firms in their daily operations.

However, TS often falls short of achieving a state-of-the-art performance as it does not explicitly take into account the value of exploration, i.e., its arm-selection rule does not consider how the DM's belief will change during the remaining time periods. This can be critical in practical settings, in which, for example, a time constraint restricts the amount of learning, or each action conveys a different amount of information, or the extra randomness in the system naturally leads to "free" exploration.

In this dissertation, we examine the use of two different DP techniques, namely, information relaxation and policy gradient, and develop two general frameworks that provide systematic ways to improve TS, with the aim of producing a better approximation of the Bayesian optimal policy.

- **Thompson sampling with information relaxation penalties** (Chapter 2). We first propose a framework that naturally generalizes TS by extending the idea of posterior sampling. An

algorithm in this framework draws a random sample of the entire future reward realizations in addition to the model parameters and decides which arm to pull by solving a deterministic reward maximization problem with respect to the sampled future scenario in the presence of penalties. We show that TS is a special case that follows from a particular penalty scheme and can be improved by incorporating penalties that reflect the value of future information more precisely.

- **Policy gradient optimization of Thompson sampling policies** (Chapter 3). We then propose a data-driven framework that can numerically optimize the control parameters of TS using the policy gradient method. While the policy gradient is a general tool for optimizing a randomized policy, it fails to work for TS since the likelihood of an arm being selected by TS cannot be written in a closed form in general. To overcome this issue, we interpret the sampled model parameters as a pseudo-action taken by TS, whose probability distribution is available in a closed form, and then apply the policy gradient in this pseudo-action space.

Comparison of above two frameworks shows that the one with information relaxation improves the performance of TS by investing additional online computation cost without need of extra control parameters, whereas the one with policy gradient does it by investing additional offline computation cost without need of application-specific analysis. Generally speaking, the former is more analytical and more suitable for situations where there exist some restrictions in the DM's decision making, whereas the latter is more practical and more widely applicable. Both frameworks leverage ideas developed in simulation literature and provide systematic ways to improve TS that achieve a more precise exploration–exploitation trade-off. We also provide theoretical analyses and numerical experiments showing that our suggested methodologies effectively fix the shortcomings of TS and achieve state-of-the-art performance in various settings.

1.2 Risk-sensitive optimal control problem via a conditional value-at-risk measure

In real-world applications, the DM often wants to be conservative in the face of uncertainties, by optimizing performance in adverse scenarios rather than focusing on average performance. This has long been an important topic in operations research and has been studied in the areas of risk-sensitive optimization and robust optimization.

In Chapter 4, we consider the use of conditional value-at-risk (CVaR) as an objective, which measures the average cost in a certain fraction of worst scenarios, i.e., the conditional average in the tail of the cost distribution. CVaR is particularly favored in practice also in theory because it offers a very intuitive quantification of uncertainty and also has nice mathematical properties. In the studies of optimal control under a risk measure, however, it has been considered difficult in general to apply the conventional DP framework due to the time-inconsistency of the risk measure (roughly speaking, a composition of CVaR measures is not a CVaR measure).

As an alternative, we leverage the idea of state augmentation that introduces an extra state variable representing the quantile value (at which the CVaR value of the future cost is measured), and develop a CVaR dynamic programming framework in the continuous-time setting. More specifically, we show that a certain type of CVaR-optimal control problem can be described as a continuous-time stochastic game between the DM who controls the original state process and an adversary who controls the quantile value process, based on the dual representation of the CVaR measure. We further derive a Bellman-like optimality equation that has a form of minimax optimization by exploiting the martingale representation theorem.

We adopt the suggested methodology to solve a "risk-sensitive optimal execution problem", given a task of liquidating a specific amount of a financial asset, the DM controls the liquidation rate adaptively to the price change so as to minimize the CVaR value of the total transaction cost, measured at a target quantile level. By solving partial differential equations that follow from the optimality equation, we derive the optimal dynamic trading strategy in a closed form, and characterize its "aggressiveness-in-the-money" behavior formally. An analytic comparison with

the optimized static trading strategy is also provided.

Chapter 2: Thompson Sampling with Information Relaxation Penalties

2.1 Introduction

Dating back to the earliest work [2, 1], multi-armed bandit (MAB) problems have been considered within a Bayesian framework, in which the unknown parameters are modeled as random variables drawn from a known prior distribution. In this setting, the problem can be viewed as a Markov decision process (MDP) with a state that is an information state describing the beliefs of unknown parameters that evolve stochastically upon each play of an arm according to Bayes' rule.

Under the objective of expected performance, where the expectation is taken with respect to the prior distribution over unknown parameters, the (Bayesian) optimal policy (OPT) is characterized by Bellman equations immediately following from the MDP formulation. In the discounted infinite-horizon setting, the celebrated Gittins index [1] determines an optimal policy, despite the fact that its computation is still challenging. In the non-discounted finite-horizon setting, which we consider, the problem becomes more difficult [3], and except for some special cases, the Bellman equations are neither analytically nor numerically tractable, due to the curse of dimensionality. In this paper, we focus on the Bayesian setting, and attempt to apply ideas from dynamic programming (DP) to develop tractable policies with good performance.

To this end, we apply the idea of *information relaxation* [4], a technique that provides a systematic way of obtaining the performance bounds on the optimal policy. In multi-period stochastic DP problems, admissible policies are required to make decisions based only on previously revealed information. The idea of information relaxation is to consider non-anticipativity as a constraint imposed on the policy space that can be relaxed, while simultaneously introducing a penalty for this relaxation into the objective, as in the usual Lagrangian relaxations of convex duality theory. Under such a relaxation, the decision maker (DM) is allowed to access future information and is asked

to solve an optimization problem so as to maximize her total reward, in the presence of penalties that punish any violation of the non-anticipativity constraint. When the penalties satisfy a condition (dual feasibility, formally defined in §2.3), the expected value of the maximal reward adjusted by the penalties provides an upper bound on the expected performance of the (non-anticipating) optimal policy.

The idea of relaxing the non-anticipativity constraint has been studied in different contexts [5, 6, 7, 8], and was later formulated as a formal framework by [4], upon which our methodology is developed. This framework has been applied to a variety of applications including optimal stopping problems [9]; linear-quadratic and linear-convex control [10, 11]; dynamic portfolio execution [12]; and more [e.g., 13, 14]. Typically, the application of this method to a specific class of MDPs requires custom analysis. In particular, it is not always easy to determine penalty functions that (1) yield a relaxation that is tractable to solve, and (2) provide tight upper bounds on the performance of the optimal policy. Moreover, the established information relaxation theory focuses on upper bounds and provides no guidance on the development of tractable policies.

Our contribution is to apply the information relaxation techniques to the finite-horizon stochastic MAB problem, explicitly exploiting the structure of a Bayesian learning process. In particular,

1. we propose a series of information relaxations and penalties of increasing computational complexity;
2. we systematically obtain the upper bounds on the best achievable expected performance that trade off between tightness and computational complexity;
3. and we develop associated (randomized) policies that generalize Thompson sampling (TS) in the finite-horizon setting.

In our framework, which we call *information relaxation sampling*, each of the penalty functions (and information relaxations) determines one policy and one performance bound given a particular problem instance specified by the time horizon and the prior beliefs. As a base case for our algorithms, we have TS [15] and the conventional regret benchmark that has been used for Bayesian regret analysis since [16]. At the other extreme, the optimal policy OPT and its expected

performance follow from the "ideal" penalty (which, not surprisingly, is intractable to compute). By picking increasingly strict information penalties, we can improve the policy and the associated bound between the two extremes of TS and OPT.

As an example, one of our algorithms, IRS.FH, is a very simple modification of TS that naturally incorporates time horizon T. Recalling that TS makes a decision based on sampled parameters for each arm from the posterior distribution in each epoch, observe that knowledge of the parameters is essentially (assuming Bayesian consistency) as informative as having an infinite number of future reward observations from each arm. By contrast, IRS.FH makes a decision based on future Bayesian estimates, updated with only $T - 1$ future reward realizations for each arm, where the rewards are sampled based on the inital posterior belief. When $T = 1$ (equivalently, at the last decision epoch), such a policy takes a myopically best action based only on the current estimates, which is indeed an optimal decision, whereas TS would still explore unnecessarily. While keeping the recursive structure of the sequential decision-making process of TS, IRS.FH naturally performs less exploration than TS as the remaining time horizon diminishes. This mitigates a common practical criticism of TS: it explores too much.

Beyond this, we propose other algorithms that more explicitly quantify the benefit of exploration and more explicitly trade off between exploration and exploitation, at the cost of additional computational complexity. As we increase the complexity, we achieve policies that improve performance, and separately provide tighter tractable computational upper bounds on the expected performance of any policy for a particular problem instance. By providing natural generalizations of TS, our work provides both a deeper understanding of TS and improved policies that do not require tuning. Since TS has been shown to be asymptotically regret optimal in some settings, e.g., by the metric of growth-rate [17] or by the metric of worst-case regret [18, 19], our improvements can at best be (asymptotically) constant factor improvements by that metric. On the other hand, TS is extremely popular in practice, and we demonstrate in numerical examples that the improvements can be significant and are likely to be of practical interest.

Moreover, we develop upper bounds on performance that are useful in their own right. Suppose

that a decision maker faces a particular problem instance and is considering any particular MAB policy (be it one we suggest or otherwise). By simulating the policy, we can find a lower bound on the performance of the optimal policy. We introduce a series of upper bounds that can also be evaluated in any problem instance via simulation. Paired with the lower bound, these provide a computational, simulation-based "confidence interval" that can be helpful to the decision maker. For example, if the upper bound and lower bound are close, the suboptimality gap of the policy under consideration is guaranteed to be small, and it is not worth investing in better policies.

2.2 Problem

2.2.1 Bayesian MAB with independent arms

We consider a Bayesian MAB problem with K *independent arms* and a *finite time horizon* T. More specifically, we define an MAB instance with a tuple $\left(K, T, \mathcal{R}, \Theta, \mathcal{P}, \mathcal{Y}, \mathbf{y}\right)$ as follows. In each period $t = 1, \ldots, T$, the decision maker (DM) selects one among K arms, each of which yields a stochastic reward whenever selected. We let $\mathcal{A} \triangleq \{1, \ldots, K\}$ denote the set of arms, and let $R_{a,n}$ denote the random variable that represents the reward from the n^{th} pull[1] of arm $a \in \mathcal{A}$. For each arm a, the rewards $\{R_{a,n}\}_{n\in\mathbb{N}}$ are independent and identically distributed according to the distribution $\mathcal{R}_a(\theta_a)$, where $\theta_a \in \Theta_a$ is the *parameter* associated with arm a:

$$R_{a,n} \sim \mathcal{R}_a(\theta_a), \quad \forall n \in \mathbb{N}, \quad \forall a \in \mathcal{A}. \tag{2.1}$$

The parameter θ_a is unknown to the DM, and is modeled as a random variable for which we have a family of *conjugate priors* $\{\mathcal{P}_a(y_a)\}_{y_a\in\mathcal{Y}_a}$, i.e., a space of distributions for θ_a that is closed under a Bayesian update with a reward realization $R_{a,n}$. Given a *hyperparameter* $y_a \in \mathcal{Y}_a$ (also called a *belief*), consider a probability measure $\mathbb{P}_{y_a}[\cdot]$ under which the parameter θ_a follows the *prior*

[1] One may consider an alternative stochastic model for the reward realization process in which the rewards are defined through a time index (e.g., $R_{a,t}$ denotes the reward from arm a in period t). This would be mathematically equivalent from the perspective of the DM. However, once the information set is relaxed, such a model is *not* equivalent to ours: in our model, the DM is not allowed to skip any future reward realizations, and this is crucial for some of the algorithms suggested in this paper. See the discussion in §2.3.3.

distribution $\mathcal{P}_a(y_a)$:

$$\theta_a \sim \mathcal{P}_a(y_a), \quad \forall a \in \mathcal{A}. \tag{2.2}$$

Let $\mathbb{E}_{y_a}[\cdot]$ denote the expected value under this probability measure. For brevity, denote the vector of parameters and hyperparameters across arms by $\boldsymbol{\theta} \triangleq (\theta_1, \ldots, \theta_K)$ and $\mathbf{y} \triangleq (y_1, \ldots, y_K)$, respectively. Define $\mathcal{R}$, Θ, $\mathcal{P}$, $\mathcal{Y}$, $\mathbb{P}_{\mathbf{y}}$, and $\mathbb{E}_{\mathbf{y}}$ analogously. We will often describe an MAB instance only with a tuple $(T, \mathbf{y})$ when the other components are clear in context.

Throughout the paper, we assume that the rewards are absolutely integrable for each hyperparameter $y_a \in \mathcal{Y}_a$:

$$\mathbb{E}_{y_a}\left[|R_{a,1}|\right] < \infty, \quad \forall y_a \in \mathcal{Y}_a,\ a \in \mathcal{A}, \tag{2.3}$$

where the expectation is taken with respect to the random realization of the parameter θ_a and also with respect to the random realization of the reward $R_{a,1}$.

We further define the *outcome* $\omega \in \Omega$ (also referred to as the future or scenario) as a combination of the parameters and all future reward realizations, i.e.,

$$\omega \triangleq \left(\boldsymbol{\theta}, (R_{a,n})_{a\in\mathcal{A},n\in\mathbb{N}}\right) \ \sim \ \mathcal{I}(\mathbf{y}), \tag{2.4}$$

that encodes all the uncertainties that the DM encounters in the environment and whose distribution is denoted by $\mathcal{I}(\mathbf{y})$.

Policy. Given an outcome ω, the reward at time t can be represented as a function of the DM's action sequence $\mathbf{a}_{1:t} = (a_1, \ldots, a_t) \in \mathcal{A}^t$, i.e.,

$$r_t(\mathbf{a}_{1:t}, \omega) \triangleq R_{a_t, n_t(\mathbf{a}_{1:t}, a_t)}, \tag{2.5}$$

where $n_t(\mathbf{a}_{1:t}, a) \triangleq \sum_{s=1}^{t} \mathbf{1}\{a_s = a\}$ counts how many times an arm a has been played up to time t (inclusive). Consequently, we define the *history* $H_t(\mathbf{a}_{1:t}, \omega)$ as the information revealed to the

DM up to time t when taking an action sequence $\mathbf{a}_{1:t}$ given the outcome ω:

$$H_t(\mathbf{a}_{1:t}, \omega) \triangleq \big(a_1, r_1(a_1, \omega), a_2, r_2(\mathbf{a}_{1:2}, \omega), \ldots, a_t, r_t(\mathbf{a}_{1:t}, \omega)\big). \tag{2.6}$$

Let $\mathbf{A}^{\pi}_{1:t}$ be the action sequence taken under the DM's policy π. We can define the natural filtration $\mathbb{F} \triangleq (\mathcal{F}_t)_{t=0,1,\ldots,T}$ where $\mathcal{F}_t \triangleq \sigma\left(H_t(\mathbf{A}^{\pi}_{1:t}, \omega)\right)$ is the σ-field generated by the history H_t.

A policy π is called *non-anticipating* if every action A^{π}_t is measurable with respect to $\mathcal{F}_{t-1}$; i.e., each decision is made based only on the information revealed prior to that time. We denote by $\Pi_{\mathbb{F}}$ the set of all non-anticipating policies, including randomized ones. The (Bayesian) *performance* of a policy π is measured by the total reward that π earns on average, i.e.,

$$V(\pi, T, \mathbf{y}) \triangleq \mathbb{E}_{\mathbf{y}}\left[\sum_{t=1}^{T} r_t(\mathbf{A}^{\pi}_{1:t}, \omega)\right], \tag{2.7}$$

where T and $\mathbf{y}$ specify, respectively, the length of the time horizon and the prior hyperparameters of given the MAB instance.

Bayesian update. Whenever the DM observes a reward realization, as a Bayesian learner, she can update her belief associated with the selected arm according to Bayes' rule. More formally, we introduce a *Bayesian update function* $\mathcal{U}_a : \mathcal{Y}_a \times \mathbb{R} \to \mathcal{Y}_a$ so that after observing a reward $r \in \mathbb{R}$ from an arm $a \in \mathcal{A}$, the hyperparameter associated with arm a is updated from y_a to $\mathcal{U}_a(y_a, r)$ (e.g., if $\theta_a \sim \mathcal{P}_a(y_a)$, then $\theta_a | R_{a,1} \sim \mathcal{P}_a(\mathcal{U}_a(y_a, R_{a,1}))$). We will often use $\mathcal{U} : \mathcal{Y} \times \mathcal{A} \times \mathbb{R} \to \mathcal{Y}$ to denote the updating of the hyperparameter vector $\mathbf{y}$; i.e., after observing a reward realization r from an arm a, the hyperparameter vector is updated from $\mathbf{y}$ to $\mathcal{U}(\mathbf{y}, a, r)$, where only the a^{th} component is updated.

We further describe the time evolution of the DM's belief throughout the decision making process. Given an outcome ω and an action sequence $\mathbf{a}_{1:t}$, the posterior hyperparameter vector at time t can be recursively expressed as

$$\mathbf{y}_t(\mathbf{a}_{1:t}, \omega; \mathbf{y}) \triangleq \mathcal{U}\left(\mathbf{y}_{t-1}(\mathbf{a}_{1:t-1}, \omega; \mathbf{y}), a_t, r_t(\mathbf{a}_{1:t}, \omega)\right), \quad \forall t \geq 1, \tag{2.8}$$

with $\mathbf{y}_0 \triangleq \mathbf{y}$. We often write $[\mathbf{y}_t(\mathbf{a}_{1:t}, \omega; \mathbf{y})]_a$ to denote the a^{th} component of $\mathbf{y}_t(\mathbf{a}_{1:t}, \omega; \mathbf{y})$. This hyperparameter vector $\mathbf{y}_t(\mathbf{a}_{1:t}, \omega; \mathbf{y})$ sufficiently describes the DM's belief given the history $H_t(\mathbf{a}_{1:t}, \omega)$.

Mean reward. We introduce several notions of mean reward that play a crucial role throughout the paper. For each arm $a \in \mathcal{A}$, we let $\mu_a(\theta_a)$ denote the *conditional mean reward* given the parameter θ_a, and let $\bar{\mu}_a(y_a)$ be the *predictive mean reward* given the hyperparameter y_a:

$$\mu_a(\theta_a) \triangleq \mathbb{E}\left[R_{a,n} \middle| \theta_a\right], \quad \bar{\mu}_a(y_a) \triangleq \mathbb{E}_{y_a}\left[\mu_a(\theta_a)\right]. \tag{2.9}$$

We further define the *posterior predictive mean reward process* $\left\{\hat{\mu}_{a,n}\right\}_{n\geq 0}$ by

$$\hat{\mu}_{a,n}(\omega; y_a) \triangleq \mathbb{E}_{y_a}\left[\mu_a(\theta_a) | R_{a,1}, \ldots, R_{a,n}\right], \tag{2.10}$$

which represents the predictive mean reward (i.e., the finite-sample Bayesian estimate of $\mu_a(\theta_a)$) after observing first n rewards associated with the arm a.

Remark 2.2.1. *Fix an arm $a \in \mathcal{A}$. The posterior predictive mean reward process $\{\hat{\mu}_{a,n}\}_{n\geq 0}$ is a martingale adapted to the filtration generated by the sequence of rewards $(R_{a,1}, R_{a,2}, R_{a,3}, \ldots)$. Furthermore, it starts at the value of the prior predictive mean reward $\bar{\mu}_a(y_a)$ and converges to the conditional mean reward $\mu_a(\theta_a)$; i.e., $\hat{\mu}_{a,0}(\omega; y_a) = \bar{\mu}_a(y_a)$ and $\lim_{n\to\infty} \hat{\mu}_{a,n}(\omega; y_a) = \mu_a(\theta_a)$ almost surely (see Proposition A.4.2 in the Appendix).*

2.2.2 Natural exponential family

We will often consider the case where the reward distribution $\mathcal{R}_a(\theta_a)$ belongs to the *natural exponential family*. In this case, the closed-form expressions are available for the aforementioned notation. For any given $\theta_a \in \Theta_a \subseteq \mathbb{R}$, the probability measure for a random reward $R_{a,n}$ is determined by

$$\mathbb{P}\left[R_{a,n} \in dr \mid \theta_a\right] = h_a(dr) \exp\left(\theta_a r - A_a(\theta_a)\right), \tag{2.11}$$

where $h_a(dr)$ is the *reference measure* and $A_a(\cdot)$ is the *log-partition function* that is a logarithm of the normalization factor. We then have a family of conjugate priors $\{\mathcal{P}_a(y_a)\}_{y_a \in \mathcal{Y}_a}$ where $\mathcal{Y}_a \triangleq \{y_a = (\xi_a, \nu_a) | \xi_a \in \mathbb{R}, \nu > 0\}$, so, for any given hyperparameter $y_a \in \mathcal{Y}_a$, the corresponding prior $\mathcal{P}_a(y_a)$ is also an exponential family distribution and can be described as

$$\mathbb{P}_{(\xi_a, \nu_a)} [\theta_a \in d\theta] = f_a(\xi_a, \nu_a) \exp(\xi_a \theta - \nu_a A_a(\theta)) \, d\theta, \tag{2.12}$$

where $f_a(\xi_a, \nu_a)$ is the normalization factor and ν_a represents the effective number of observations. Within this family of conjugate priors, it is well known that the posterior distribution can be expressed as

$$\mathbb{P}_{(\xi_a, \nu_a)} \left[\theta_a \in d\theta \,\middle|\, R_{a,1}, \ldots, R_{a,n}\right] = \mathbb{P}_{\left(\xi_a + \sum_{i=1}^n R_{a,i},\ \nu_a + n\right)} [\theta_a \in d\theta] \,. \tag{2.13}$$

This property can also be expressed via the Bayesian update function as $\mathcal{U}_a((\xi_a, \nu_a), r) = (\xi_a + r, \nu_a + 1)$. We also have the following identities for the mean reward metrics:

$$\mu_a(\theta_a) = A_a'(\theta_a), \quad \bar{\mu}_a(\xi_a, \nu_a) = \frac{\xi_a}{\nu_a}, \quad \hat{\mu}_{a,n}(\omega; \xi_a, \nu_a) = \frac{\xi_a + \sum_{i=1}^n R_{a,n}}{\nu_a + n}, \tag{2.14}$$

where $A_a' \triangleq dA_a / d\theta_a$. We refer the reader to [20] for further details.

Bernoulli and Gaussian MABs. We briefly illustrate the Bernoulli MAB and Gaussian MAB as representative examples of the problem instance described by a natural exponential family. In the Bernoulli MAB, the rewards of an arm are Bernoulli random variables whose success probability is drawn from a Beta distribution. In the Gaussian MAB, the rewards of an arm are normally distributed with an unknown mean and a known noise variance where the mean is also normally distributed. Table 2.1 summarizes the previously defined notation.

	Bernoulli MAB	Gaussian MAB
Prior distribution	$\mu_a \sim \text{Beta}(\alpha_a, \beta_a)$	$\mu_a \sim \mathcal{N}(m_a, v_a^2)$
Reward distribution	$R_{a,n} \sim \text{Bernoulli}(\mu_a)$	$R_{a,n} \sim \mathcal{N}(\mu_a, \sigma_a^2)$
Parameter θ_a	$\theta_a = \log \frac{\mu_a}{1-\mu_a}$	$\theta_a = \frac{\mu_a}{\sigma_a^2}$
Hyperparameters ξ_a, ν_a	$\xi_a = \alpha_a$, $\nu_a = \alpha_a + \beta_a$	$\xi_a = \frac{m_a \sigma_a^2}{v_a^2}$, $\nu_a = \frac{\sigma_a^2}{v_a^2}$
Reference measure h_a	$h_a(dr) = \delta_0(dr) + \delta_1(dr)$	$h_a(dr) = \frac{1}{\sqrt{2\pi\sigma_a^2}} \exp\left(-\frac{r^2}{\sigma_a^2}\right) dr$
Log-partition function A_a	$A_a(\theta_a) = \log\left(1 + e^{\theta_a}\right)$	$A_a(\theta_a) = \frac{\sigma_a^2 \theta_a^2}{2}$
Mean reward μ_a	$\mu_a(\theta_a) = \frac{e^{\theta_a}}{1+e^{\theta_a}}$	$\mu_a(\theta_a) = \sigma_a^2 \theta_a$
Predictive mean $\bar{\mu}_a$	$\bar{\mu}_a(\alpha_a, \beta_a) = \frac{\alpha_a}{\alpha_a + \beta_a}$	$\bar{\mu}_a(m_a, v_a^2) = m_a$

Table 2.1: Description of a Bernoulli MAB and a Gaussian MAB. Here, $\delta_x(dr)$ denotes a Dirac measure that has a single atom at x.

2.2.3 Bayesian optimal policy

In a Bayesian framework, the MAB problem can be viewed as a Markov decision process (MDP) in which a state corresponds to an information state (or belief state) of the DM. It has the following recursive structure that we will exploit throughout the paper. Given an MAB instance with time horizon T and prior belief $\mathbf{y}$, suppose that the DM has just earned r by pulling an arm a at time $t = 1$. Then the remaining problem for the DM is equivalent to an MAB instance with time horizon $T - 1$ and prior belief $\mathcal{U}(\mathbf{y}, a, r)$. Based on this Markovian structure, we obtain the following Bellman equations for the MAB problem: for all $T \in \mathbb{N}$ and $\mathbf{y} \in \mathcal{Y}$,

$$Q^*(T, \mathbf{y}, a) \triangleq \mathbb{E}_{\mathbf{y}}\left[R_{a,1} + V^*(T-1, \mathcal{U}(\mathbf{y}, a, R_{a,1}))\right], \tag{2.15}$$

$$V^*(T, \mathbf{y}) \triangleq \max_{a \in \mathcal{A}} Q^*(T-1, \mathbf{y}, a), \tag{2.16}$$

with $V^*(0, \mathbf{y}) \triangleq 0$ for all $\mathbf{y} \in \mathcal{Y}$. The value function $V^*(T, \mathbf{y})$ represents the best possible performance that a non-anticipating policy can achieve in the MAB problem specified by the time

horizon T and the prior belief $\mathbf{y}$, or equivalently, the maximum expected future reward that one can earn during T remaining periods[2] when the current belief is $\mathbf{y}$.

While Bellman equations are, in general, intractable to solve and directly apply, they offer a characterization of the *Bayesian optimal policy* (OPT). At a certain moment, when the remaining time horizon is T and the belief is $\mathbf{y}$, OPT takes an action with the largest state-action value (Q-value), i.e., pulls the arm $A^* = \operatorname{argmax}_a Q^*(T, \mathbf{y}, a)$, and this action selection procedure is repeated while updating T and $\mathbf{y}$ according to Bayes' rule as described in Algorithm 1. Such a policy achieves the best possible performance among all non-anticipating policies:

$$V^*(T, \mathbf{y}) = \sup_{\pi \in \Pi_{\mathbb{F}}} V(\pi, T, \mathbf{y}) = V(\text{OPT}, T, \mathbf{y}), \quad \forall T \in \mathbb{N}, \mathbf{y} \in \mathcal{Y}. \tag{2.17}$$

Algorithm 1: Bayesian optimal policy (OPT)

Function `OPT` $(T, \mathbf{y})$

```
// T:remaining time horizon, y:current belief
```

return $\operatorname{argmax}_a Q^*(T, \mathbf{y}, a)$

Procedure `OPT-Outer` $(T, \mathbf{y})$

```
// T:time horizon, y:prior belief
```

$\mathbf{y}_0 \leftarrow \mathbf{y}$

for $t = 1, 2, \ldots, T$ **do**

  Select $A_t \leftarrow$ `OPT` $(T - t + 1, \mathbf{y}_{t-1})$

  Earn and observe a reward r_t and update belief $\mathbf{y}_t \leftarrow \mathcal{U}(\mathbf{y}_{t-1}, A_t, r_t)$

end

[2]We intentionally refrain from indexing the value function V^* by time t, since such a representation conceals the Markovian structure of the Bayesian MAB problem and leads to complicated expressions for the variables that exploit this Markovian structure. To avoid confusion, the horizon T will be written as an argument to functions whereas the time index t will be written as a subscript, throughout the paper.

2.2.4 Thompson sampling

Thompson sampling (TS) is a simple heuristic that makes decisions based on random sampling. When the remaining time is T and the current belief is $\mathbf{y}$, it samples the parameters $\tilde{\boldsymbol{\theta}}$ from the prior[3] distribution at that moment, $\mathcal{P}(\mathbf{y})$, and pulls the arm that is believed to be best given the sampled parameters $\tilde{\boldsymbol{\theta}}$, i.e., takes action $A^{\mathrm{TS}} = \mathrm{argmax}_a \, \mu_a(\tilde{\theta}_a)$. Like OPT, it repeats this procedure at every decision epoch while updating the belief $\mathbf{y}$ whenever a reward realization is observed.

Algorithm 2: Arm selection rule of Thompson sampling when remaining time is T and current belief is $\mathbf{y}$

Function `TS` (T, $\mathbf{y}$)

`// T:remaining time horizon, y:current belief`

1 Sample parameters $\tilde{\boldsymbol{\theta}} \sim \mathcal{P}(\mathbf{y})$

2 **return** $\mathrm{argmax}_a\{\mu_a(\tilde{\theta}_a)\}$

Note that TS does not take into account the time information when making a decision. It applies the identical sampling and selection rule, irrespective of the remaining time periods. This often leads to the unnecessary explorations near the end of the horizon, which motivates our framework.

2.3 Information Relaxation Sampling

We apply the information relaxation framework [4] to the Bayesian MAB problem and propose a general framework which we call *information relaxation sampling* (IRS). The main idea behind the information relaxation is to relax the information constraint so that the decision maker (DM) is allowed to exploit some future information that is supposed to be unknown. As in the usual Lagrangian relaxation, an upper bound on the best possible performance can be obtained by solving the relaxed problem.

To motivate in detail, let us consider a situation under which the parameters $\boldsymbol{\theta}$ are revealed to

[3] Conventionally, the term "posterior distribution" is used to describe the distribution that TS samples the parameters from. We explicitly use "prior distribution" instead: for example, at time $t = 1$, the parameters are apparently sampled from the prior, not the posterior, distribution. After observing a reward realization, we will have a posterior but it will become a prior at the next decision epoch.

the DM when the remaining period is T and the current belief is $\mathbf{y}$. The optimal action for this DM is to keep playing the arm with the highest mean reward, i.e., $\operatorname{argmax}_a \mu_a(\theta_a)$, and by doing so will earn $\mathbb{E}_{\mathbf{y}}[T \times \max_a \mu_a(\theta_a)]$ on average, which is indeed an upper bound on the performance of the optimal policy, $V^*(T, \mathbf{y})$.

Let us now postulate a situation under which the same kind of DM is informed with sampled parameters $\tilde{\boldsymbol{\theta}}$ that are drawn from the distribution $\mathcal{P}(\mathbf{y})$. For this (falsely informed) DM, the optimal action is again to play the arm with the highest mean reward but now with respect to the sampled parameters, i.e., $\operatorname{argmax}_a \mu_a(\tilde{\theta}_a)$. This procedure effectively describes the arm selection rule of Thompson sampling in the situation specified by the remaining horizon T and the current belief $\mathbf{y}$.

Above, we motivated a performance bound, $\mathbb{E}_{\mathbf{y}}[T \times \max_a \mu_a(\theta_a)]$, and a non-anticipating policy, TS, from the relaxation of the parameter information. Analogously, we can produce another performance bound and another policy by considering a different set of future information to relax: the performance bound is obtained by computing how much the clairvoyant DM can earn with this additional information; and the policy is obtained by speculating which action the same kind of DM will take if the additional information is replaced with sampled (simulated) instance.

We will particularly consider the relaxations of information that are less effective than the full parameter information for the DM to maximize her future payoff. This will result in tighter relaxations, in the sense of a better (tighter) performance upper bound as well as a better performing policy.

In what follows, we formalize this idea utilizing the notion of *information relaxation penalties* that allows us to describe and control the benefit from having additional information explicitly. We will first describe the general framework and then propose a specific family of penalties that are particularly suitable for Bayesian MAB problems.

Information relaxation penalties and the inner problem. Applying the information relaxation framework developed by [4], we relax the non-anticipativity constraint imposed on policy space

$\Pi_{\mathbb{F}}$ (i.e., A_t^π is $\mathcal{F}_{t-1}$-measurable). Without loss of generality,[4] we consider the perfect information relaxation under which the DM is allowed to first observe all future outcomes in advance, and then pick an action (i.e., A_t^π is $\sigma(\omega)$-measurable). As in any other Lagrangian relaxation, we impose penalties on the DM for violating the non-anticipativity constraint.

We introduce a *penalty function* $z_t(\mathbf{a}_{1:t}, \omega; T, \mathbf{y})$ to denote the penalty that the DM incurs at time t, when taking an action sequence $\mathbf{a}_{1:t}$ given an outcome ω for an MAB problem with time horizon T and prior belief $\mathbf{y}$. The clairvoyant DM can find the best action sequence that is optimal for this particular outcome ω in the presence of penalties z_t, by solving the following (deterministic) optimization problem, referred to as the *inner problem*:

$$\underset{\mathbf{a}_{1:T} \in \mathcal{A}^T}{\text{maximize}} \quad \sum_{t=1}^{T} r_t(\mathbf{a}_{1:t}, \omega) - z_t(\mathbf{a}_{1:t}, \omega; T, \mathbf{y}). \tag{$*$}$$

Definition 2.3.1 (Dual feasibility)**.** *Given T and $\mathbf{y}$, a penalty function z_t is dual feasible if it is a zero mean for any non-anticipating policy $\pi \in \Pi_{\mathbb{F}}$, i.e.,*

$$\mathbb{E}_{\mathbf{y}}\left[\sum_{t=1}^{T} z_t(\mathbf{A}_{1:t}^\pi, \omega; T, \mathbf{y})\right] = 0, \quad \forall \pi \in \Pi_{\mathbb{F}}. \tag{2.18}$$

We remark that the mapping $\mathbf{a}_{1:t} \mapsto z_t(\mathbf{a}_{1:t}, \omega)$ is a stochastic function of the action sequence $\mathbf{a}_{1:t}$ since the outcome ω is random. This dual feasibility condition requires that the DM who makes decisions on the natural filtration will receive zero penalties in expectation.

The complexity of the inner problem depends very much on the penalty function. Assuming that the penalty function can be evaluated in $O(1)$ computation, an enumerative brute-force optimization of the inner problem may require $O(K^T)$ computations. In what follows, we will illustrate that for suitably designed penalty functions, the inner problem exhibits a recursive structure and thus can be solved effectively using dynamic programming techniques.

[4] Any partial information relaxation can be equivalently described within the perfect information relaxation by adding additional terms into the penalty function. See the discussion after Theorem 2.3.1.

IRS performance bound. We let $W^z(T, \mathbf{y})$ be the expected maximal value of the inner problem $(*)$, when the outcome ω is randomly drawn from its prior distribution $\mathcal{I}(\mathbf{y})$, i.e., the expected total payoff that a clairvoyant DM can achieve in the presence of penalties.:

$$W^z(T, \mathbf{y}) \triangleq \mathbb{E}_{\mathbf{y}} \left[\max_{\mathbf{a}_{1:T} \in \mathcal{A}^T} \left\{ \sum_{t=1}^{T} r_t(\mathbf{a}_{1:t}, \omega) - z_t(\mathbf{a}_{1:t}, \omega; T, \mathbf{y}) \right\} \right]. \tag{2.19}$$

Once we have an algorithm to solve the inner problem, this value can be computed numerically via simulation: let $\omega_1, \omega_2, \ldots, \omega_S$ be the samples independently drawn from $\mathcal{I}(\mathbf{y})$, and W_s be the the maximal value of the inner problem with respect to ω_s for each $s = 1, \ldots, S$ separately. The bound W^z can be computed by taking the average of these maximal values, i.e., $\frac{1}{S} \sum_{s=1}^{S} W_s$. The following theorem shows that W^z is indeed a valid performance bound of the stochastic MAB problem.

Theorem 2.3.1 (Weak duality and strong duality)**.** *If the penalty function z_t is dual feasible, W^z is an upper bound on the optimal value V^*:*

$$\textit{(Weak duality)} \qquad W^z(T, \mathbf{y}) \geq V^*(T, \mathbf{y}). \tag{2.20}$$

There exists a dual feasible penalty function denoted by z_t^{ideal}, such that

$$\textit{(Strong duality)} \qquad W^{\text{ideal}}(T, \mathbf{y}) = V^*(T, \mathbf{y}). \tag{2.21}$$

The ideal penalty function z_t^{ideal} has the following functional form:

$$\begin{aligned} z_t^{\text{ideal}}(\mathbf{a}_{1:t}, \omega; T, \mathbf{y}) \triangleq\ & r_t(\mathbf{a}_{1:t}, \omega) - \mathbb{E}_{\mathbf{y}} \left[r_t(\mathbf{a}_{1:t}, \omega) \,|H_{t-1}(\mathbf{a}_{1:t-1}, \omega) \right] \\ & + V^*\left(T - t, \mathbf{y}_t(\mathbf{a}_{1:t}, \omega; \mathbf{y})\right) - \mathbb{E}_{\mathbf{y}} \left[V^*\left(T - t, \mathbf{y}_t(\mathbf{a}_{1:t}, \omega; \mathbf{y})\right) | \, H_{t-1}(\mathbf{a}_{1:t-1}, \omega) \right]. \end{aligned} \tag{2.22}$$

Recall that a dual feasible penalty function does not penalize (in expectation) non-anticipating policies, which include OPT. Even when the future information is available, the DM can earn V^* under the penalties by implementing OPT without taking advantage of future information. When

the DM makes use of future information, she can always outperform OPT, which leads to the weak duality result. The ideal penalty z_t^{ideal} precisely penalizes for the additional profit extracted from using the future information, thereby removing any incentive to deviate from OPT and resulting in the strong duality.

The ideal penalty is, of course, intractable, but its structure highlights what a good penalty may look like. It implies that there are two sources of additional profit: in DP terminology, one from knowing future immediate rewards and one from knowing future state transitions, each of which will be taken into account later in this paper.

As another implication, it also shows that relaxing more the available information can always be compensated by adding associated terms to the penalty function. That is, a partial information relaxation (e.g., A_t^π is measurable w.r.t. $\mathcal{G}_{t-1}$ such that $\sigma(H_{t-1}) \subseteq \mathcal{G}_{t-1} \subseteq \sigma(\omega)$) with some penalty function $z_t^{\mathbb{G}}$ is equivalent to the perfect information relaxation (i.e., A_t^π is measurable w.r.t. $\sigma(\omega)$) with a penalty function $z_t^{\mathbb{G}} + z_t^{\sigma(\omega)\setminus\mathbb{G}}$ if the additional term $z_t^{\sigma(\omega)\setminus\mathbb{G}}$ exactly penalizes the relative benefit from having more information $\sigma(\omega)$ than $\mathcal{G}_{t-1}$. Hence, it is sufficient to consider the perfect information relaxation, as we do in this paper, and the actual amount of information available for the DM can be equivalently controlled by adjusting the penalty function.

Before proceeding, we remark that the above results are already well established in [4] (see Lemma 2.1 and Theorem 2.3 therein) for a general class of MDP problems, except for a subtle difference regarding the assumption on the predictability of reward realizations. In MDP problems, the reward at each state is typically assumed to be deterministic (otherwise, it is replaced with its expected value), since the stochastic evolution of the state is of a major concern. By contrast, in MAB problems it is essential to consider the randomness of rewards since learning from the noisy reward realizations is of a major concern, and therefore, we do not assume that r_t is measurable with respect to $\sigma(H_{t-1})$. As a consequence, our ideal penalty function (2.22) has a slightly different functional form than the one formulated in [4].[5] We further exploit this fact when designing a variety of penalty functions.

[5] [4] show that $z_t^{\text{ideal}} = V^*(T-t, \mathbf{y}_t) - \mathbb{E}[V^*(T-t, \mathbf{y}_t) | H_{t-1}]$, when r_t is assumed to be measurable with respect to $\sigma(H_{t-1})$ and so $r_t - \mathbb{E}[r_t | H_{t-1}] = 0$.

IRS policy. Since the true outcome ω is not available in reality, it cannot be used in online decision making. We derive a non-anticipating policy by leveraging the idea of "posterior sampling," which utilizes the sampled outcome $\tilde{\omega}$ instead of the true outcome ω.

Given a penalty function z_t, we characterize a randomized and non-anticipating IRS policy π^z as follows. Exploiting the recursive structure of a Bayesian MAB problem, the policy π^z specifies "which arm to pull when the remaining time is T and the current belief is $\mathbf{y}$," i.e., the very first action that it would take in an MAB instance with horizon T and prior belief $\mathbf{y}$. Given T and $\mathbf{y}$, it (i) first randomly generates an outcome $\tilde{\omega}$ (i.e., sampling from $\mathcal{I}(\mathbf{y})$), (ii) solves the inner problem to find a best action sequence $\tilde{\mathbf{a}}^*_{1:T}$ with respect to this randomly generated outcome $\tilde{\omega}$ in the presence of penalties z_t, and (iii) takes the first action $\tilde{a}^*_1$ that the clairvoyant optimal solution $\tilde{\mathbf{a}}^*_{1:T}$ suggests. Analogous to TS and OPT, **it repeats steps (i)–(iii) at every decision epoch**, while

updating the remaining time T and belief $\mathbf{y}$ upon each decision making and reward realization.

Algorithm 3: Information relaxation sampling (IRS) policy

Function `IRS` $(T, \mathbf{y}; z)$

`// T:remaining time horizon, y:current belief`

1 Sample an outcome $\tilde{\omega} \sim \mathcal{I}(\mathbf{y})$: Equivalently, for each $a \in \mathcal{A}$,

$$\tilde{\theta}_a \sim \mathcal{P}_a(y_a), \quad \tilde{R}_{a,n} \sim \mathcal{R}_a(\tilde{\theta}_a), \quad \forall n \in \{1, \dots, T\}.$$

2 Find the best action sequence with respect to the sampled outcome $\tilde{\omega}$ under penalties z_t:

$$\tilde{\mathbf{a}}^*_{1:T} \leftarrow \operatorname*{argmax}_{\mathbf{a}_{1:T} \in \mathcal{A}^T} \left\{ \sum_{s=1}^{T} r_s(\mathbf{a}_{1:t}, \tilde{\omega}) - z_s(\mathbf{a}_{1:s}, \tilde{\omega}; T, \mathbf{y}) \right\}.$$

3 **return** $\tilde{a}^*_1$

Procedure `IRS-Outer` $(T, \mathbf{y}; z)$

`// T:time horizon, y:prior belief`

1 $\mathbf{y}_0 \leftarrow \mathbf{y}$

2 **for** $t = 1, 2, \dots, T$ **do**

3 Pull $A_t \leftarrow$ `IRS` $(T - t + 1, \mathbf{y}_{t-1}; z)$

4 Earn and observe a reward r_t and update belief $\mathbf{y}_t \leftarrow \mathcal{U}(\mathbf{y}_{t-1}, A_t, r_t)$

end

In step (i), the random generation of the outcome $\tilde{\omega}$ given the belief $\mathbf{y}$ is equivalent to, for each arm $a \in \mathcal{A}$, sampling the parameter from its posterior, $\tilde{\theta}_a \sim \mathcal{P}_a(y_a)$, and then sampling the future reward realizations, $\tilde{R}_{a,n} \sim \mathcal{R}_a(\tilde{\theta}_a)$ for $n = 1, \dots, T$. In other words, the IRS policy π^z randomly generates (simulates) a plausible future scenario within its own probability space specified by T and $\mathbf{y}$.

The optimization problem in the step (ii) is identical to the inner problem $(*)$ except that the true outcome ω is replaced with the sampled one $\tilde{\omega}$. Therefore, the dynamic programming algorithm that solves the inner problem can also be utilized for this online decision-making process, not only

for the computation of performance bound W^z. Note that there can be multiple solutions to this optimization problem and the tie-breaking rule may affect the performance of the policy. We do not observe that the choice of tie-breaking rule is significance in our numerical experiments. In some instances that follow, however, we will adopt a specific tie-breaking rule for the purpose of theoretical analysis.

Also note that in step (iii) only the first action $\tilde{a}_1^*$ of the optimal solution $\tilde{\mathbf{a}}_{1:T}^*$ is utilized, and at the following decision epoch a new outcome is sampled based on the updated belief. If we consider an MAB instance with time horizon T, the policy π^z solves T different instances of the inner problem throughout the entire decision-making process, with a decreasing length of time horizon, from T to 1, and with a stochastically evolving belief state. See the IRS-OUTER procedure in Algorithm 3, which is in fact identical to that employed in OPT and TS.

Remark 2.3.1. *The ideal penalty yields the Bayesian optimal policy, i.e.,* $\pi^{\text{ideal}} = \text{OPT}$.

Recall that the ideal penalty (2.22) yields the performance bound W^{ideal} that is equal to the best achievable performance V^*, because the DM under the ideal penalty has no incentive to utilize any future information. For the same reason, the corresponding IRS policy π^{ideal} does not utilize the (randomly generated) future information in its decision making, and tries to make the best decision based only on the information revealed so far. Therefore, its decision should always coincide with the Bayesian optimal policy's decision.

Choice of penalty functions. We have so far described the general framework that takes a penalty function z_t as input, and yields a performance bound W^z and a policy π^z as outputs. While any dual feasible penalty functions can be utilized in general, we propose the following set of penalty functions that are particularly suitable for the MAB problems:

$$z_t^{\mathrm{TS}}(\mathbf{a}_{1:t}, \omega) \triangleq r_t(\mathbf{a}_{1:t}, \omega) - \mathbb{E}\left[r_t(\mathbf{a}_{1:t}, \omega) \,|\boldsymbol{\theta}\right], \tag{2.23}$$

$$z_t^{\text{IRS.FH}}(\mathbf{a}_{1:t}, \omega) \triangleq r_t(\mathbf{a}_{1:t}, \omega) - \mathbb{E}_{\mathbf{y}}\left[r_t(\mathbf{a}_{1:t}, \omega) \,|\hat{\boldsymbol{\mu}}_{T-1}(\omega)\right], \tag{2.24}$$

$$z_t^{\text{IRS.V-ZERO}}(\mathbf{a}_{1:t}, \omega) \triangleq r_t(\mathbf{a}_{1:t}, \omega) - \mathbb{E}_{\mathbf{y}}\left[r_t(\mathbf{a}_{1:t}, \omega) \,|H_{t-1}(\mathbf{a}_{1:t-1}, \omega)\right], \tag{2.25}$$

$$\begin{aligned} z_t^{\text{IRS.V-EMAX}}(\mathbf{a}_{1:t}, \omega) \triangleq\ & r_t(\mathbf{a}_{1:t}, \omega) - \mathbb{E}_{\mathbf{y}}\left[r_t(\mathbf{a}_{1:t}, \omega) \,|H_{t-1}(\mathbf{a}_{1:t-1}, \omega)\right] \\ & + W^{\mathrm{TS}}\left(T-t, \mathbf{y}_t(\mathbf{a}_{1:t}, \omega)\right) - \mathbb{E}_{\mathbf{y}}\left[W^{\mathrm{TS}}\left(T-t, \mathbf{y}_t(\mathbf{a}_{1:t}, \omega)\right) \middle| H_{t-1}(\mathbf{a}_{1:t-1}, \omega)\right], \end{aligned} \tag{2.26}$$

where $\hat{\boldsymbol{\mu}}_{T-1}(\omega; \mathbf{y}) \triangleq \left(\hat{\mu}_{a,T-1}(\omega; y_a)\right)_{a \in \mathcal{A}}$ and the dependency of some expressions on T and $\mathbf{y}$ is suppressed for clarity. Also recall that the ideal penalty is given by

$$\begin{aligned} z_t^{\text{ideal}}(\mathbf{a}_{1:t}, \omega) \triangleq\ & r_t(\mathbf{a}_{1:t}, \omega) - \mathbb{E}_{\mathbf{y}}\left[r_t(\mathbf{a}_{1:t}, \omega) \,|H_{t-1}(\mathbf{a}_{1:t-1}, \omega)\right] \\ & + V^*\left(T-t, \mathbf{y}_t(\mathbf{a}_{1:t}, \omega)\right) - \mathbb{E}_{\mathbf{y}}\left[V^*\left(T-t, \mathbf{y}_t(\mathbf{a}_{1:t}, \omega)\right) \middle| H_{t-1}(\mathbf{a}_{1:t-1}, \omega)\right]. \end{aligned} \tag{2.27}$$

We can show that these penalty functions satisfy the dual feasibility condition (Definition 2.3.1); see Appendix A.3.3 for a formal proof.

Remark 2.3.2. *All penalty functions* (2.23)–(2.27) *are dual feasible.*

This set of penalty functions results in a set of policies that ranges from Thompson sampling (TS) to the Bayesian optimal policy (OPT) and a set of performance bounds that ranges from the conventional regret benchmark W^{TS} ($= \mathbb{E}[T \times \max_a \mu_a(\theta_a)]$) to the optimal value function W^{ideal} ($= V^*$). More specifically, at one extreme, the simplest penalty function z_t^{TS} yields TS and W^{TS} as outputs, and at the other extreme, the ideal penalty function z_t^{ideal} yields OPT and V^* which would be optimal. The other three penalty functions ($z_t^{\text{IRS.FH}}$, $z_t^{\text{IRS.V-ZERO}}$, and $z_t^{\text{IRS.V-EMAX}}$) connect the two extremes and are sequentially "better", where we informally say that a penalty function is better than another if it is closer to the ideal penalty function and thus yields a better performing policy and a tighter performance bound. Deferring detailed explanations to §2.3.1–§2.3.4, we briefly illustrate general principles to design "good" penalty functions and motivate these penalty

functions.

In design of information relaxation penalties, we first need to determine to which information set we relax the non-anticipativity constraint, i.e., what kind of additional information will be revealed to the DM in the relaxation. Although we have described our framework based on the perfect information relaxation (i.e., the relaxation in which the DM perfectly knows the entire future outcomes ω), any imperfect information relaxation can be equivalently described within the perfect information relaxation using a properly constructed penalty function.[6] Among the suggested penalty functions,[7] z_t^{TS} is the one that corresponds to the information relaxation to the parameter information $\boldsymbol{\theta}$, $z_t^{\mathrm{IRS.FH}}$ corresponds to the information relaxation to the posterior predictive mean rewards $\hat{\boldsymbol{\mu}}_{T-1}$ (i.e., the finite-sample mean-reward estimates), and z_t^{ideal} corresponds to no information relaxation.

One principle to motivate a better penalty function is to choose a smaller set of future information for the relaxation. When less additional information is revealed to the DM in the relaxation, the additional profit that the DM can extract from this information becomes smaller, and hence the DM has to make more realistic decisions that rely more on the currently available information rather than the future information that is supposed to be unknown. Comparing $z_t^{\mathrm{IRS.FH}}$ with z_t^{TS}, one may notice that the finite-sample mean-reward estimates $\hat{\boldsymbol{\mu}}_{T-1}$ are less informative than the parameters $\boldsymbol{\theta}$ for the DM to exploit in her profit maximization because, in terms of mean-reward estimation, the parameters are informative as much as an infinite number of observations (i.e., $\mathbb{E}[\mu_a(\theta_a)|\boldsymbol{\theta}] = \lim_{T\to\infty} \mathbb{E}[\mu_a(\theta_a)|R_{a,1},\ldots,R_{a,T-1}] = \lim_{T\to\infty} \hat{\mu}_{a,T-1}$). In this sense, $z_t^{\mathrm{IRS.FH}}$ is better than z_t^{TS}, and resulting policy $\pi^{\mathrm{IRS.FH}}$ and performance bound $W^{\mathrm{IRS.FH}}$ improve upon TS and W^{TS} toward OPT and V^*.

Another principle to motivate a better penalty function is to adopt a more precise approximation of the ideal penalty function z_t^{ideal}, particularly regarding the terms containing the optimal

[6] In fact, this is the main idea underlying the existence of the ideal penalty function; see the discussion after Theorem 2.3.1.

[7] We can motivate one more penalty function that corresponds to the perfect information relaxation. Such a penalty function is simply given by $z_t \equiv 0$, which is illustrated in Appendix A.1. However, we do not suggest it's use since it is even worse than z_t^{TS}.

value function V^*. In the presence of penalties that reflect the value of the additional information more accurately, the DM has less incentive to exploit this additional information in the relaxed decision making problem, and similarly to the above argument, this leads to more realistic decisions. Among our suggestions, $z_t^{\text{IRS.V-ZERO}}$ approximates the term V^* with zero, and $z_t^{\text{IRS.V-EMAX}}$ approximates the term V^* with a tractable upper bound W^{TS}. By doing so, $z_t^{\text{IRS.V-EMAX}}$ takes into account the continuation value of each action explicitly and improves upon $z_t^{\text{IRS.V-ZERO}}$.

Consider the inner problem associated with each choice of penalty function (2.23)–(2.27). Recall that each inner problem is a deterministic multi-period decision making problem that has a form of $\max_{\mathbf{a}_{1:T} \in \mathcal{A}^T} \sum_{t=1}^{T} r_t(\mathbf{a}_{1:t}) - z_t(\mathbf{a}_{1:t})$. A penalty function z_t effectively redefines what the DM earns at each time, i.e., $r_t(\mathbf{a}_{1:t})$ is replaced with $r_t(\mathbf{a}_{1:t}) - z_t(\mathbf{a}_{1:t})$. More specifically, the penalty function z_t^{TS} effectively replaces the realized rewards associated with each arm with their expected value given parameters $\boldsymbol{\theta}$; as does $z_t^{\text{IRS.FH}}$ (with their expected value given the finite-sample mean-reward estimates $\hat{\boldsymbol{\mu}}_{T-1}$); as does $z_t^{\text{IRS.V-ZERO}}$ (with their expected value conditional on how many times the arm has previously been selected up to each point in time).

Penalty function	Policy	Performance bound	Inner problem	Run time
z_t^{TS}	TS	W^{TS}	Find a best arm given parameters.	$O(K)$
$z_t^{\text{IRS.FH}}$	$\pi^{\text{IRS.FH}}$	$W^{\text{IRS.FH}}$	Find a best arm given finite observations.	$O(K)$ or $O(KT)$
$z_t^{\text{IRS.V-ZERO}}$	$\pi^{\text{IRS.V-ZERO}}$	$W^{\text{IRS.V-ZERO}}$	Find an optimal allocation of T pulls.	$O(KT^2)$
$z_t^{\text{IRS.V-EMAX}}$	$\pi^{\text{IRS.V-EMAX}}$	$W^{\text{IRS.V-EMAX}}$	Find an optimal action sequence.	$O(KT^K)$
z_t^{ideal}	OPT	V^*	Solve Bellman equations.	–

Table 2.2: List of algorithms following from penalty functions (2.23)–(2.27). TS refers to Thompson sampling and OPT refers to the Bayesian optimal policy. Run time represents the computational complexity of solving one instance of the inner problem $(*)$, that is, the time required to obtain one sample in a computation of performance bound W^z or to decide which arm to select in each period in a run of policy π^z.

Table 2.2 summarizes these inner problems. As we sequentially increase the computational

complexity of a penalty function, from z_t^{TS} to z_t^{ideal}, the penalty function more accurately penalizes the benefit from knowing future outcomes, i.e., more explicitly prevents the DM from exploiting future information. As a result, the inner problem becomes closer to the original stochastic optimization problem, which results in a better performing policy and a tighter performance bound. Using this approach, we achieve a family of algorithms that are intuitive and tractable, exhibiting a trade-off between quality and computational efficiency. See Appendix §A.1 for an illustrative example.

The run time in Table 2.2 represents the computational complexity of solving one instance of the inner problem, i.e., the time it takes to obtain one sample in a computation of performance bound W^z or to decide which arm to select in each period in a run of policy π^z. In this run-time analysis, performing the Bayesian belief updating and the sampling of a random variable is counted as a single operation.

2.3.1 Thompson sampling revisited

With the penalty function $z_t^{\mathrm{TS}}(\mathbf{a}_{1:t}, \omega) \triangleq r_t(\mathbf{a}_{1:t}, \omega) - \mu_{a_t}(\theta_{a_t})$, the inner problem $(*)$ reduces to

$$\max_{\mathbf{a}_{1:T} \in \mathcal{A}^T} \left\{ \sum_{t=1}^{T} r_t(\mathbf{a}_{1:t}, \omega) - z_t(\mathbf{a}_{1:t}, \omega) \right\} = \max_{\mathbf{a}_{1:T} \in \mathcal{A}^T} \left\{ \sum_{t=1}^{T} \mu_{a_t}(\theta_{a_t}) \right\} = T \times \max_{a \in \mathcal{A}} \mu_a(\theta_a). \tag{2.28}$$

Given an outcome ω, and in the presence of penalties, a hindsight optimal action sequence is to keep pulling the true best arm, i.e., $\operatorname{argmax}_a \mu_a(\theta_a)$, for T times in a row. The resulting performance bound W^{TS} reduces to the conventional performance benchmark,

$$W^{\mathrm{TS}}(T, \mathbf{y}) = \mathbb{E}_{\mathbf{y}} \left[T \times \max_{a \in \mathcal{A}} \mu_a(\theta_a) \right], \tag{2.29}$$

which measures how much the DM could have achieved if the parameters had been revealed in advance.

Remark 2.3.3. *The performance bound W^{TS} is the conventional benchmark that has been widely*

used in the Bayesian regret analysis [16, 21, 22]. The Bayesian regret of a policy π is defined as

$$\mathit{BayesRegret}(\pi, T, \mathbf{y}) \triangleq \mathbb{E}_{\mathbf{y}}\left[\sum_{t=1}^{T} \max_{a} \mu_a(\theta_a) - \mu_{A_t^\pi}(\theta_{A_t^\pi})\right] = W^{\mathrm{TS}}(T, \mathbf{y}) - V(\pi, T, \mathbf{y}), \tag{2.30}$$

which quantifies the suboptimality of the policy π.

It is trivial to see that the corresponding policy π^{TS} is equivalent to Thompson sampling. The policy π^{TS} utilizes a sampled outcome $\tilde{\omega}$ instead of the true outcome ω; accordingly, it selects an arm $A^{\mathrm{TS}} = \operatorname{argmax}_a \mu_a(\tilde{\theta}_a)$, where $\tilde{\boldsymbol{\theta}} \sim \mathcal{P}(\mathbf{y})$, which is identical to the procedure described in Algorithm 2. In order for the policy π^{TS} to make a decision at a certain time, note that it does not need to sample future rewards, and thus it requires $O(K)$ computations only.

2.3.2 IRS.FH

Recall that $\hat{\mu}_{a,T-1}(\omega; y_a)$ is the posterior predictive mean reward of an arm a that the DM will have after observing $T-1$ reward realizations $R_{a,1}, \ldots, R_{a,T-1}$ given the initial belief y_a:

$$\hat{\mu}_{a,T-1}(\omega; y_a) \triangleq \mathbb{E}_{y_a}\left[\mu_a(\theta_a) \,\middle|\, R_{a,1}, \ldots, R_{a,T-1}\right]. \tag{2.31}$$

Given (2.24), the optimal solution to the inner problem $(*)$ is to always pull the arm with the highest posterior predictive mean reward, i.e., $\operatorname{argmax}_a \hat{\mu}_{a,T-1}(\omega; y_a)$:

$$\max_{\mathbf{a}_{1:T} \in \mathcal{A}^T} \left\{\sum_{t=1}^{T} r_t(\mathbf{a}_{1:t}, \omega) - z_t^{\mathrm{IRS.FH}}(\mathbf{a}_{1:t}, \omega)\right\} = \max_{\mathbf{a}_{1:T} \in \mathcal{A}^T} \left\{\sum_{t=1}^{T} \hat{\mu}_{a_t,T-1}(\omega)\right\} = T \times \max_{a \in \mathcal{A}} \hat{\mu}_{a,T-1}(\omega). \tag{2.32}$$

This inner problem yields the performance bound $W^{\mathrm{IRS.FH}}$, such that

$$W^{\mathrm{IRS.FH}}(T, \mathbf{y}) = \mathbb{E}_{\mathbf{y}}\left[T \times \max_{a \in \mathcal{A}} \hat{\mu}_{a,T-1}(\omega; y_a)\right], \tag{2.33}$$

and the policy $\pi^{\text{IRS.FH}}$ that is implemented in Algorithm 4.

Algorithm 4: Arm selection rule of $\pi^{\text{IRS.FH}}$ when remaining time is T and current belief is $\mathbf{y}$

Function `IRS.FH(`$T, \mathbf{y}$`)`

`// T:remaining time horizon, y:current belief`

1 Sample parameters $\tilde{\boldsymbol{\theta}} \sim \mathcal{P}(\mathbf{y})$ and rewards $\tilde{R}_{a,n} \sim \mathcal{R}_a(\tilde{\theta}_a)$, $\forall n \in \{1, \dots, T\}$, $\forall a \in \mathcal{A}$.

2 **return** $\operatorname{argmax}_a \left\{ \mathbb{E}_{y_a} \left[\mu_a(\theta_a) \middle| R_{a,1} = \tilde{R}_{a,1}, \dots, R_{a,T-1} = \tilde{R}_{a,T-1} \right] \right\}$

IRS.FH (FH stands for finite horizon) is almost identical to TS except that the conditional mean reward $\mu_a(\theta_a)$ is replaced with the posterior predictive mean reward $\hat{\mu}_{a,T-1}(\omega)$. As a finite-sample Bayesian estimate of the conditional mean reward, $\hat{\mu}_{a,T-1}(\omega)$ is less informative than $\mu_a(\theta_a)$ from the DM's perspective. In terms of mean reward estimation, the DM will never be able to identify $\mu_a(\theta_a)$ perfectly within a finite horizon, i.e., knowing the parameters is equivalent to having an infinite number of observations. The inner problem of TS requires the DM to "identify the best arm based on an infinite number of samples," whereas that of IRS.FH requires the DM to "identify the best arm based on a finite number of samples" and takes into account the length of the time horizon explicitly. By restricting the DM's access to fewer information, IRS.FH requires the DM to be more realistic, that is, to consider the uncertainties more precisely.

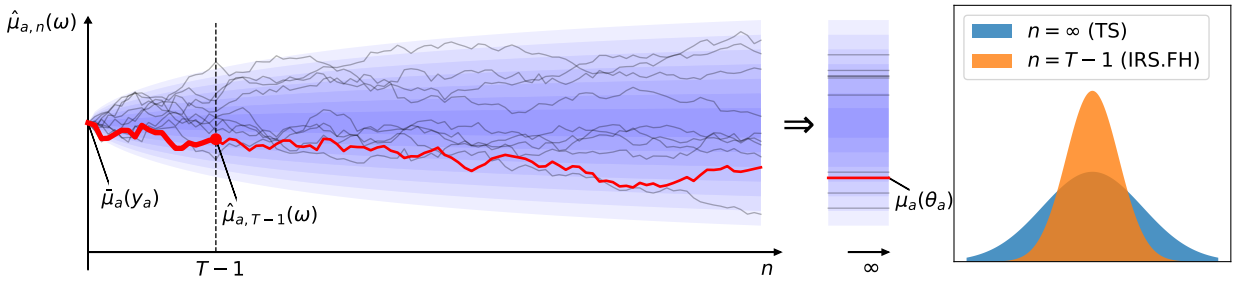

Figure 2.1: (Left) Sample paths of posterior predictive mean reward process of an arm a, $\{\hat{\mu}_{a,n}(\omega)\}_{n \geq 0}$. This process is a martingale that starts at (prior) predictive mean $\bar{\mu}_a$ and converges to conditional mean μ_a (Remark 2.2.1). (Right) The distributions of $\hat{\mu}_{a,T-1}$ and μ_a: $\hat{\mu}_{a,T-1}$ is more concentrated than μ_a, while all have the same mean $\bar{\mu}_a(y_a)$.

To sharpen our comparison between IRS.FH and TS, let us compare the variance of $\hat{\mu}_{a,T-1}(\omega)$

and $\mu_a(\theta_a)$ induced by the randomness of outcome ω. As depicted in Figure 2.1, μ_a is more widely distributed than $\hat{\mu}_{a,T-1}$ because a larger (infinite vs. $T-1$) number of samples makes it easier for the posterior to deviate from the initial prior (see also Remark 2.2.1). By Jensen's inequality, we further have $W^{\text{IRS.FH}} = \mathbb{E}[T \times \max_a \hat{\mu}_{a,T-1}(\omega)] \leq \mathbb{E}[T \times \max_a \mu_a(\theta_a)] = W^{\text{TS}}$ for any problem instance, meaning that IRS.FH yields a performance bound that is tighter than the conventional benchmark. Also note that the same argument holds for the comparison between $\hat{\mu}_{a,T-1}(\tilde{\omega})$ and $\mu_a(\tilde{\theta}_a)$ since the synthesized outcome $\tilde{\omega}$ is identically distributed with the (true) outcome ω. The variability of $\hat{\mu}_{a,T-1}(\tilde{\omega})$ (respectively, $\mu_a(\tilde{\theta}_a)$) governs the randomness of the action taken by policy $\pi^{\text{IRS.FH}}$ (resp., π^{TS}), i.e., $A^{\text{IRS.FH}} = \operatorname{argmax}_a \hat{\mu}_{a,T-1}(\tilde{\omega})$ (resp., $A^{\text{TS}} = \operatorname{argmax}_a \mu_a(\tilde{\theta}_a)$). Given T and $\mathbf{y}$, the policy $\pi^{\text{IRS.FH}}$ performs fewer random explorations than TS, as it is less likely to deviate from the myopic decision to play an arm with the largest current estimate $\bar{\mu}_a(y_a)$. More desirably, the degree of exploration of $\pi^{\text{IRS.FH}}$ is controlled by the remaining time horizon as the variance of $\hat{\mu}_{a,T-1}(\omega)$ depends on T. At the last decision epoch ($T = 1$), $\pi^{\text{IRS.FH}}$ takes a myopic action that is indeed optimal.

Efficiently sampling $\hat{\mu}_{a,T-1}(\tilde{\omega})$ for natural exponential families. In order to obtain $\hat{\mu}_{a,T-1}(\tilde{\omega})$ for each arm a for a synthesized outcome $\tilde{\omega}$, one may apply Bayes' rule sequentially for each reward realization, which will take $O(T)$ computations per arm.

As discussed in §2.2.2, in an MAB were the reward distribution is a natural exponential family, the posterior predictive mean reward is given by

$$\hat{\mu}_{a,T-1}(\tilde{\omega}; \xi_a, \nu_a) = \frac{\xi_a + \sum_{n=1}^{T-1} \tilde{R}_{a,n}}{\nu_a + T - 1}. \tag{2.34}$$

Therefore, it is sufficient to sample the sum of $T-1$ future rewards, $\sum_{n=1}^{T-1} \tilde{R}_{a,n}$, in otder to sample the posterior predictive mean reward. Observe that the conditional distribution of the sum given $\tilde{\theta}_a$ also belongs to the natural exponential family, induced by a log-partition function $(T-1)A_a(\tilde{\theta}_a)$. This distribution may be tractable to compute: for example, its distribution is Binomial$(T-1, \mu_a(\tilde{\theta}_a))$ for the Beta-Bernoulli case, and $\mathcal{N}\big((T-1)\cdot\mu_a(\tilde{\theta}_a), (T-1)\cdot\sigma_a^2\big)$ for the Gaussian case. In these

settings, we can sample the sum $\sum_{n=1}^{T-1} \tilde{R}_{a,n}$ directly from the tractable distribution (after sampling $\tilde{\theta}_a$) using $O(1)$ computation, and then use it to compute $\hat{\mu}_{a,T-1}(\tilde{\omega})$ without sequentially updating the belief. In such cases, a single decision of $\pi^{\text{IRS.FH}}$ can be made within $O(K)$ operations, independent of T, similar in computational complexity to TS.

2.3.3 IRS.V-ZERO

IRS.V-ZERO introduces a further complication in that its inner problem requires the DM to consider her causal process in the course of solving the inner problem. Under the penalty $z_t^{\text{IRS.V-ZERO}}$ given in (2.25), the DM at time t earns $\mathbb{E}\left[r_t(\mathbf{a}_{1:t}, \omega) \,|H_{t-1}(\mathbf{a}_{1:t-1}, \omega)\right]$, the expected mean reward that she can infer from observations prior to time t. As we defined $R_{a,n}$ to be a reward from the n^{th} pull on arm a (not the pull at time n), the posterior belief associated with each arm is determined only by the number of past pulls performed on that arm. Recall that $\hat{\mu}_{a,n}(\omega)$ is the expected mean reward of arm a that the DM can infer from the first n reward realizations:

$$\hat{\mu}_{a,n}(\omega; y_a) \triangleq \mathbb{E}_{y_a}\left[\mu_a(\theta_a) \,\middle|R_{a,1}, \ldots, R_{a,n}\right]. \tag{2.35}$$

Therefore, the DM earns $\hat{\mu}_{a,n-1}(\omega)$ from the n^{th} pull on arm a, irrespective of the detailed sequence of the past actions. More formally, the DM's earning at time t is

$$r_t(\mathbf{a}_{1:t}, \omega) - z_t^{\text{IRS.V-ZERO}}(\mathbf{a}_{1:t}, \omega) = \mathbb{E}_{\mathbf{y}}\left[\mu_{a_t}(\theta_{a_t}) \,|H_{t-1}(\mathbf{a}_{1:t-1}, \omega)\right] = \hat{\mu}_{a_t, n_{t-1}(\mathbf{a}_{1:t-1}, a_t)}(\omega), \tag{2.36}$$

where $n_{t-1}(\mathbf{a}_{1:t-1}, a)$, defined in (2.5), denotes the number of pulls conducted on a particular arm a prior to time t.

Let $S_{a,n}(\omega) \triangleq \sum_{i=1}^{n} \hat{\mu}_{a,i-1}(\omega)$ be the cumulative payoff from the first n pulls of an arm a. Given an outcome ω, we observe that the total payoff is determined only by the total number of pulls on each arm, and not the sequence in which the arms have been pulled. Therefore, solving the inner problem $(*)$ is equivalent to "finding the optimal allocation $(n_1^*, n_2^*, \ldots, n_K^*)$ among T remaining

opportunities": more formally,

$$\max_{\mathbf{a}_{1:T}\in\mathcal{A}^T}\left\{\sum_{t=1}^{T}\hat{\mu}_{a_t,n_{t-1}(\mathbf{a}_{1:t-1},a_t)}\right\}=\max_{\mathbf{a}_{1:T}\in\mathcal{A}^T}\left\{\sum_{a=1}^{K}\sum_{n=1}^{n_T(\mathbf{a}_{1:T},a)}\hat{\mu}_{a,n-1}\right\}=\max_{\mathbf{n}_{1:K}\in N_T}\left\{\sum_{a=1}^{K}S_{a,n_a}\right\},\qquad(2.37)$$

where $N_T \triangleq \{(n_1,\dots,n_K)\in\mathbb{N}_0^K : \sum_{a=1}^K n_a = T\}$ is the set of all feasible allocations. Once the $S_{a,n}$'s are computed, we can solve this inner problem within $O(KT^2)$ operations by sequentially applying the sup convolution K times. The detailed implementation is provided in §A.2.1.

The policy $\pi^{\text{IRS.V-ZERO}}$ further needs to decide which arm to pull given the optimal allocation $(\tilde{n}_1^*,\tilde{n}_2^*,\dots,\tilde{n}_K^*)$ that is obtained for the sampled outcome $\tilde{\omega}$. In principle, any arm a that was included in the solution of the inner problem, $\tilde{n}_a^* > 0$, would suffice, but we suggest a selection rule by which the arm that needs the most pulls is chosen, i.e., $A^{\text{IRS.V-ZERO}} = \text{argmax}_a\, \tilde{n}_a^*$. This guarantees that $\pi^{\text{IRS.V-ZERO}}$ behaves like TS when T is large, as formally stated in Proposition 2.4.1.

Comparison with TS **and** IRS.FH. Recall that in the inner problems of TS and IRS.FH, the DM at time t earns $\mathbb{E}[r_t|\boldsymbol{\theta}]$ and $\mathbb{E}[r_t|\hat{\boldsymbol{\mu}}_{T-1}]$, respectively, which are the mean reward estimates that rely on the information not available at the moment; e.g., $\hat{\mu}_{a,T-1}$ is revealed only after playing the arm a for $T-1$ times. IRS.V-ZERO is more restrictive for the DM in the sense that she at time t earns $\mathbb{E}[r_t|H_{t-1}]$, which does not include any information that does not belong to H_{t-1}. IRS.V-ZERO reflects the fact that the n^{th} reward of an arm will not be revealed unless the arm is pulled n times, and its inner problem requires the DM to allocate a pull in order to incorporate the next reward realization into her information set; thus learning about an arm comes at the cost of sacrificing an opportunity to learn about the other arms.

More specifically, let us focus on the total payoff of a particular allocation $(n_1,\dots,n_K)$ under each penalty function $z_t^{\text{IRS.V-ZERO}}$ and $z_t^{\text{IRS.FH}}$. The allocation yields $\sum_{a=1}^K S_{a,n_a}(\omega)$ in the inner problem of IRS.V-ZERO whereas the same allocation yields $\sum_{a=1}^K n_a \times \hat{\mu}_{a,T-1}(\omega)$ in the inner problem of IRS.FH. In terms of variability originating from the randomness of ω, we observe that each summand $S_{a,n_a}(\omega) = \sum_{i=1}^{n_a}\hat{\mu}_{a,i-1}(\omega)$ is less noisy than its counterpart $n_a \times \hat{\mu}_{a,T-1}(\omega)$ since

a larger number of observations makes it easier for the posterior to deviate from the initial prior and hence the variance of individual terms $\hat{\mu}_{a,0}(\omega), \ldots, \hat{\mu}_{a,n_a-1}(\omega)$ is smaller than the variance of $\hat{\mu}_{a,T-1}(\omega)$ and, therefore, $\sum_{a=1}^{K} S_{a,n_a}(\omega)$ is smaller than $\sum_{a=1}^{K} n_a \times \hat{\mu}_{a,T-1}(\omega)$. Analogous to the comparison between IRS.FH and TS, we have that IRS.V-ZERO yields a performance bound $W^{\text{IRS.V-ZERO}}$ that is tighter than $W^{\text{IRS.FH}}$ (formally stated in Theorem 2.4.1) and a policy $\pi^{\text{IRS.V-ZERO}}$ that performs fewer random explorations than $\pi^{\text{IRS.FH}}$.

2.3.4 IRS.V-EMAX

Under perfect information relaxation, the DM perfectly knows not only (i) what she will earn at future times but also (ii) how her belief will evolve as a result of her action sequence. The previous algorithms focus on the former component by making the DM adjust the future rewards by conditioning (e.g., $\mathbb{E}[r_t(a_t)|\boldsymbol{\theta}]$, $\mathbb{E}[r_t(a_t)|\hat{\boldsymbol{\mu}}_{T-1}]$ and $\mathbb{E}[r_t(a_t)|H_{t-1}]$). IRS.V-EMAX also focuses on the latter component as well by charging the DM an additional cost for using the information on her future belief transitions.

To motivate this in detail, recall that the ideal penalty z_t^{ideal} (2.22) is

$$\begin{aligned} z_t^{\text{ideal}}(\mathbf{a}_{1:t}, \omega) \triangleq\ & r_t(\mathbf{a}_{1:t}, \omega) - \mathbb{E}\left[r_t(\mathbf{a}_{1:t}, \omega)\,|H_{t-1}(\mathbf{a}_{1:t-1}, \omega)\right] \\ & + V^*\left(T-t, \mathbf{y}_t(\mathbf{a}_{1:t}, \omega)\right) - \mathbb{E}\left[V^*\left(T-t, \mathbf{y}_t(\mathbf{a}_{1:t}, \omega)\right)|\, H_{t-1}(\mathbf{a}_{1:t-1}, \omega)\right], \end{aligned} \tag{2.38}$$

where $V^*(T-t, \mathbf{y}_t)$ measures the value of having a belief $\mathbf{y}_t$ at a future time $t+1$. Note that, at the moment the DM takes an action a_t, the next belief state $\mathbf{y}_t = \mathcal{U}(\mathbf{y}_{t-1}, a_t, r_t)$ is not measurable with respect to the natural filtration $\sigma(H_{t-1})$ since the next observation r_t is unknown. In DP terms, the conditional expectation $\mathbb{E}\left[V^*(T-t, \mathbf{y}_t)|\, H_{t-1}\right]$ captures the expected value of a (random) next state given the current state. Accordingly, the gap between its realized value and its expected value, $V^*(T-t, \mathbf{y}_t) - \mathbb{E}\left[V^*(T-t, \mathbf{y}_t)|\, H_{t-1}\right]$, measures the additional gain from knowing the next belief state $\mathbf{y}_t$. In addition to the term $r_t - \mathbb{E}\left[r_t\,|H_{t-1}\right]$ $(= z_t^{\text{IRS.V-ZERO}})$, which measures the benefit from knowing which action will yield a large immediate reward, the ideal penalty also penalizes the

long-term benefit from knowing which action will lead to a favorable belief state.

The penalty function $z_t^{\text{IRS.V-EMAX}}$ is obtained from z_t^{ideal} by replacing $V^*(T, \mathbf{y})$ with $W^{\text{TS}}(T, \mathbf{y})$, which is rather tractable. The use of $W^{\text{TS}}(T, \mathbf{y}) \triangleq \mathbb{E}_{\mathbf{y}}\left[T \times \max_a \mu_a(\theta_a)\right]$, introduced in (2.29), leads to a simple expression for its conditional expectation: since $\boldsymbol{\theta}|H_{t-1}$ is distributed with $\mathcal{P}(\mathbf{y}_{t-1})$, we have

$$\mathbb{E}_{\mathbf{y}}\left[W^{\text{TS}}\left(T-t, \mathbf{y}_t\right)\middle| H_{t-1}\right] = (T-t) \times \mathbb{E}_{\mathbf{y}}\left[\max_a \mu_a(\theta_a)\middle| H_{t-1}\right] \tag{2.39}$$

$$= (T-t) \times \mathbb{E}_{\mathbf{y}_{t-1}}\left[\max_a \mu_a(\theta_a)\right] \tag{2.40}$$

$$= W^{\text{TS}}\left(T-t, \mathbf{y}_{t-1}\right). \tag{2.41}$$

In the associated inner problem, the payoff that the DM earns at time t is

$$r_t(\mathbf{a}_{1:t}, \omega) - z_t^{\text{IRS.V-EMAX}}(\mathbf{a}_{1:t}, \omega) \tag{2.42}$$

$$= \hat{\mu}_{a_t, n_{t-1}(\mathbf{a}_{1:t-1}, a_t)}(\omega) - W^{\text{TS}}\left(T-t, \mathbf{y}_t(\mathbf{a}_{1:t}, \omega)\right) + W^{\text{TS}}\left(T-t, \mathbf{y}_{t-1}(\mathbf{a}_{1:t-1}, \omega)\right) \tag{2.43}$$

$$= \bar{\mu}_{a_t}\left([\mathbf{y}_{t-1}(\mathbf{a}_{1:t-1}, \omega)]_{a_t}\right) - W^{\text{TS}}\left(T-t, \mathbf{y}_t(\mathbf{a}_{1:t}, \omega)\right) + W^{\text{TS}}\left(T-t, \mathbf{y}_{t-1}(\mathbf{a}_{1:t-1}, \omega)\right), \tag{2.44}$$

which is completely determined by the prior belief $\mathbf{y}_{t-1}$ and the posterior belief $\mathbf{y}_t$.

We further observe that, given ω, the future belief $\mathbf{y}_t(\mathbf{a}_{1:t}, \omega)$ depends only on how many times each arm has been pulled, irrespective of the sequence of the pulls. For example, consider two action sequences $\mathbf{a}_{1:t}^A = (1, 1, 2, 1, 2)$ and $\mathbf{a}_{1:t}^B = (2, 1, 1, 2, 1)$. Even though the order of observations would differ, in both cases the agent would observe $(R_{1,1}, R_{1,2}, R_{1,3})$ from arm 1 and $(R_{2,1}, R_{2,2})$ from arm 2 and end up with the same belief $\mathbf{y}_t(\mathbf{a}_{1:t}^A, \omega) = \mathbf{y}_t(\mathbf{a}_{1:t}^B, \omega)$. We may conclude from this observation that a belief state can be sufficiently parameterized with the pull counts $\mathbf{n}_{1:K} = (n_1, \ldots, n_K)$ instead of action sequence $\mathbf{a}_{1:t}$.

As a result, the total number of possible future beliefs is $O(T^K)$, not $O(K^T)$, and we can come up with a dynamic programming algorithm that solves the inner problem within $O(c_W T^K + KT^K)$ computations where c_W is the cost of numerically calculating $W^{\text{TS}}(T, \mathbf{y})$. We refer the interested

reader to §A.2.2.

2.3.5 IRS.INDEX policy

Finally, we propose the IRS.INDEX policy, which does not strictly belong to the IRS framework, and does not produce a performance bound, but does exhibit strong empirical performance.

Roughly speaking, the IRS.INDEX is a single-sample approximation of the finite-horizon Gittins index [23], where the approximation is motivated by IRS.V-EMAX algorithm. It first solves the single-armed bandit problem for each arm in isolation, and makes a decision based on the results of these subproblems.

Single-armed bandit problem. Consider a special case of an MAB instance in which there is a single arm a that yields stochastic rewards $R_{a,n} \sim \mathcal{R}_a(\theta_a)$ with an outside option that yields a deterministic reward λ. We have a prior distribution $\mathcal{P}_a(y_a)$ over unknown parameter θ_a whereas the deterministic reward λ is known a priori.

Given an outcome $\omega_a = \big(\theta_a, (R_{a,n})_{n\in\mathbb{N}}\big)$, we can simulate the future belief trajectory $(y_{a,n})_{n\in\{0,\dots,T\}}$, where $y_{a,n}$ is the belief after n reward realizations are observed:

$$y_{a,0} \triangleq y_a, \quad y_{a,n} \triangleq \mathcal{U}_a(y_{a,n-1}, R_{a,n}), \quad \forall n = 1, \dots, T. \tag{2.45}$$

Let $V^*(T, y_a, \lambda)$ be the optimal value function associated with this single-armed bandit problem. We consider the penalty function $z_t^{\text{IRS.V-EMAX}}$ in which the value function $V^*(T, y_a, \lambda)$ is approximated by $W^{\text{TS}}(T, y_a, \lambda) = \mathbb{E}_{y_a}\left[T \times \max(\mu_a(\theta_a), \lambda)\right]$. We define $\mathcal{A} \triangleq \{0, 1\}$ such that $a_t = 1$ if the stochastic arm at time t is selected, and $a_t = 0$ if the outside option is selected. The associated

inner problem is

$$\text{maximize} \quad \sum_{t=1}^{T} \hat{\mu}_{a,n_t-1}(\omega_a) \cdot \mathbf{1}\{a_t = 1\} + \lambda \cdot \mathbf{1}\{a_t = 0\} - (T-t) \times \left(\Gamma^{\lambda}_{n_t}(\omega_a) - \Gamma^{\lambda}_{n_{t-1}}(\omega_a)\right) \tag{2.46}$$

$$\text{subject to} \quad n_t = \sum_{s=1}^{t} \mathbf{1}\{a_t = 1\}, \quad a_t \in \{0,1\}, \quad \forall t = 1,\ldots,T, \tag{2.47}$$

where $\hat{\mu}_{a,n}(\omega_a) \triangleq \mathbb{E}_{y_a}[\mu_a(\theta_a)|R_{a,1},\ldots,R_{a,n}] = \bar{\mu}_a(y_{a,n})$ and

$$\Gamma^{\lambda}_{n}(\omega_a) \triangleq \mathbb{E}_{y_{a,n}}\left[\max(\mu_a(\theta_a),\lambda)\right]. \tag{2.48}$$

With some algebra (Proposition A.2.1 in §A.2.3), we can reformulate the optimization problem as

$$\max_{0\le n\le T} \left\{ T \times \Gamma^{\lambda}_{0}(\omega_a) + (T-n) \times \left(\lambda - \min_{0\le i\le n} \Gamma^{\lambda}_{i}(\omega_a)\right) + \sum_{i=1}^{n} \left(\hat{\mu}_{a,i-1}(\omega_a) - \Gamma^{\lambda}_{i-1}(\omega_a)\right) \right\}, \tag{2.49}$$

where the decision variable n is the total number of pulls on the stochastic arm.

Let $\varphi_a(\lambda,\omega_a)$ be the (maximal) relative benefit from pulling the stochastic arm against not pulling at all:

$$\varphi_a(\lambda,\omega_a) \triangleq \max_{1\le n\le T} \left\{ T \times \Gamma^{\lambda}_{0} + (T-n) \times \left(\lambda - \min_{0\le i\le n} \Gamma^{\lambda}_{i}\right) + \sum_{i=1}^{n} \left(\hat{\mu}_{a,i-1} - \Gamma^{\lambda}_{i-1}\right) \right\} - T \times \lambda. \tag{2.50}$$

Note that $\max\{\cdot\}$ was taken over $n \ge 1$. We interpret the meaning of the sign of $\varphi_a(\lambda,\omega_a)$ as follows: given an outcome ω_a, the stochastic arm is worth trying against the deterministic outside option λ if $\varphi_a(\lambda,\omega_a) \ge 0$, and not worth trying if $\varphi_a(\lambda,\omega_a) < 0$.

Given ω_a and λ, the value of $\varphi_a(\lambda,\omega_a)$ can be computed in $O(T)$ operations by precalculating $\sum_{i=1}^{n} \hat{\mu}_{a,i-1}(\omega_a)$, $\min_{0\le i\le n} \Gamma^{\lambda}_{i}(\omega_a)$, and $\sum_{i=1}^{n} \Gamma^{\lambda}_{i-1}(\omega_a)$ over $n = 1,\ldots,T$ sequentially. The single-armed bandit problem has an additional advantage of computational efficiency: in contrast to the implementation of IRS.V-EMAX in the multi-arm setting, the approximate value function (captured by Γ^{λ}_{n}) often admits a closed-form expression in the single-armed setting. In the cases of the

Beta-Bernoulli MAB and the Gaussian MAB, for example, we have

$$\mathbb{E}_{\mu\sim\mathrm{Beta}(\alpha,\beta)}\left[\max\left(\mu,\lambda\right)\right] = \lambda \times F^{\mathrm{beta}}_{\alpha,\beta}\left(\lambda\right) + \frac{\alpha}{\alpha+\beta} \times \left(1 - F^{\mathrm{beta}}_{\alpha+1,\beta}\left(\lambda\right)\right), \tag{2.51}$$

$$\mathbb{E}_{\mu\sim\mathcal{N}(m,\nu^2)}\left[\max\left(\mu,\lambda\right)\right] = m + (\lambda - m) \times \Phi\left(\nu^{-1}(\lambda - m)\right) + \nu \times \phi\left(\nu^{-1}(\lambda - m)\right), \tag{2.52}$$

where $F^{\mathrm{beta}}_{\alpha,\beta}(\cdot)$ represents the c.d.f. of $\mathrm{Beta}(\alpha,\beta)$ distribution, and $\Phi(\cdot)$ and $\phi(\cdot)$ represent the c.d.f. and the p.d.f. of the standard normal distribution, respectively. With these expressions, $\Gamma^{\lambda}_{n}(\omega_a)$'s can be computed very efficiently without using numerical integration or Monte Carlo sampling.

Index policy. We now return to the original MAB problem with K arms. Recall that the single-armed bandit algorithm tells us whether an arm (given an outcome ω_a) is worth trying against the deterministic reward λ. We use this algorithm as a module to compute the index of each arm.

More specifically, consider a certain decision epoch when the remaining time is T and the belief is $\mathbf{y}$. For each arm $a = 1, \ldots, K$ separately, the policy $\pi^{\textsc{Irs.Index}}$ samples the future outcome $\tilde{\omega}_a$ (i.e., draws $\tilde{\theta}_a \sim \mathcal{P}_a(y_a)$ and $\tilde{R}_{a,n} \sim \mathcal{R}_a(\tilde{\theta}_a)$ for $n = 1, \ldots, T$), and finds a threshold value on the deterministic outside option that makes the arm barely worth trying:

$$\lambda^*_a(\tilde{\omega}_a) \triangleq \sup\left\{\lambda \in \mathbb{R}\,;\ \varphi_a(\lambda, \tilde{\omega}_a) \geq 0\right\}. \tag{2.53}$$

By the definition of $\varphi_a(\lambda, \omega_a)$, the threshold value $\lambda^*_a(\tilde{\omega}_a)$ measures the value of arm a as an opportunity cost of not pulling arm a at all, given a particular outcome $\tilde{\omega}_a$. We use the value $\lambda^*_a(\tilde{\omega}_a)$ as an index of arm a so that the index policy plays the arm with the largest index, i.e., $A^{\textsc{Irs.Index}} = \mathrm{argmax}_a\, \lambda^*_a(\tilde{\omega}_a)$.

Although the monotonicity of the mapping $\lambda \mapsto \varphi_a(\lambda, \tilde{\omega}_a)$ is not theoretically proven, we observe that the bisection search works sufficiently well in our numerical experiments. Since each instance of single-armed bandit problems requires $O(T)$ computations to solve, the entire procedure for arm selection requires a run time of $O(c_b \times KT)$, where c_b represents the number of

iterations in a bisection search. See §A.2.3 for the implementation details.

In addition to the IRS.INDEX policy described above, some numerical experiments include a heuristic variation of it, called IRS.INDEX*, that is obtained by using

$$\varphi_a(\lambda, \omega_a) \triangleq \max_{1 \le n \le T} \left\{ \sum_{i=1}^{n} \left(\hat{\mu}_{a,i-1}(\omega_a) - \lambda - \left(\Gamma_i^{\lambda}(\omega_a) - \Gamma_0^{\lambda}(\omega_a) \right) \right) \right\}, \tag{2.54}$$

instead of (2.50). This alternative formulation yields indices that are relatively stable across the different samples of outcome $\tilde{\omega}_a$.

We note that our index, $\lambda_a^*(\tilde{\omega}_a)$, is a random approximation of the finite-horizon Gittins (FH-Gittins) index studied in [23], [24], and [25]. The original FH-Gittins algorithm precisely solves the single-armed bandit problem, which is shown to be an optimal stopping problem in which one must decide when to stop pulling the stochastic arm as one's belief state evolves stochastically. Applying the information relaxation framework to the single-armed bandit problem, we solve, instead, a simple deterministic problem in which one must find a deterministic schedule optimized to a particular belief trajectory associated with a randomly generated outcome $\tilde{\omega}$. As in the previous algorithms, the penalties help us to obtain a solution close to the optimal stopping policy of the original single-armed bandit problem.

2.4 Analysis

In this section, we provide theoretical analyses that characterize IRS policies and performance bounds in particular for TS, IRS.FH, and IRS.V-ZERO.

Remark 2.4.1 (Single-period optimality)**.** *When $T = 1$, all of the policies $\pi^{\text{IRS.FH}}$, $\pi^{\text{IRS.V-ZERO}}$, $\pi^{\text{IRS.V-EMAX}}$, and $\pi^{\text{IRS.INDEX}}$ take the optimal action; i.e., they pull the myopically best arm $A^* = \operatorname{argmax}_a \bar{\mu}_a(y_a)$.*

Proposition 2.4.1 (Asymptotic behavior)**.** *Assume that $\mu_i(\theta_i) \neq \mu_j(\theta_j)$ almost surely for any two distinct arms $i \neq j$. As $T \nearrow \infty$, the distribution of the $\pi^{\text{IRS.FH}}$'s action converges to that of*

Thompson sampling:

$$\lim_{T\to\infty} \mathbb{P}\left[A^{\text{IRS.FH}}(T,\mathbf{y}) = a\right] = \mathbb{P}\left[A^{\text{TS}}(\mathbf{y}) = a\right], \quad \forall a \in \mathcal{A}. \tag{2.55}$$

Similarly, so does the distribution of the $\pi^{\text{IRS.V-ZERO}}$'s action:[8]

$$\lim_{T\to\infty} \mathbb{P}\left[A^{\text{IRS.V-ZERO}}(T,\mathbf{y}) = a\right] = \mathbb{P}\left[A^{\text{TS}}(\mathbf{y}) = a\right], \quad \forall a \in \mathcal{A}. \tag{2.56}$$

$A^{\text{TS}}(\mathbf{y})$, $A^{\text{IRS.FH}}(T,\mathbf{y})$ and $A^{\text{IRS.V-ZERO}}(T,\mathbf{y})$ denote the action taken by policies π^{TS}, $\pi^{\text{IRS.FH}}$, and $\pi^{\text{IRS.V-ZERO}}$, respectively, when the remaining time is T and the current belief is $\mathbf{y}$. These actions are random variables, since each of these policies uses a randomly sampled outcome $\tilde{\omega}$ of its own. Remark 2.4.1 can be easily verified by observing that, when $T = 1$, $r_1(a,\omega) - z_1(a,\omega;T,\mathbf{y}) = \bar{\mu}_a(y_a)$ for any $a \in \mathcal{A}$ for each of the penalty functions. The results in Proposition 2.4.1 follow from Remark 2.2.1 stating that the posterior predictive mean reward process converges to the conditional mean reward, i.e., $\lim_{n\to\infty} \hat{\mu}_{a,n}(\tilde{\omega}) = \mu_a(\tilde{\theta}_a)$. The assumption $\mu_i(\theta_i) \neq \mu_j(\theta_j)$ is made to avoid the ambiguity of the tie-breaking rule that is used in TS.

Remark 2.4.1 and Proposition 2.4.1 illustrate that $\pi^{\text{IRS.FH}}$ and $\pi^{\text{IRS.V-ZERO}}$ behave like TS during the initial decision epochs, gradually shift toward the myopic scheme, and end up with the optimal decision; by contrast, TS continues to explore. The transition from exploration to exploitation under these IRS policies occurs smoothly, without relying on an auxiliary control parameter. While maintaining their recursive structure, IRS policies take into account the time horizon T, and naturally balance exploitation and exploration.

Theorem 2.4.1 (Monotonicity of performance bounds)**.** IRS.FH *and* IRS.V-ZERO *monotonically improve the performance bound*

$$W^{\text{TS}}(T,\mathbf{y}) \geq W^{\text{IRS.FH}}(T,\mathbf{y}) \geq W^{\text{IRS.V-ZERO}}(T,\mathbf{y}), \tag{2.57}$$

[8] We assume a particular selection rule such that $\tilde{a}^{\text{IRS.V-ZERO}} = \text{argmax}_a\, \tilde{n}^*_a$, as discussed in §2.3.3.

and also

$$W^{\mathrm{TS}}(T, \mathbf{y}) \geq W^{\textsc{Irs.V-EMax}}(T, \mathbf{y}). \tag{2.58}$$

Recall that $W^{\mathrm{TS}}(T, \mathbf{y}) = \mathbb{E}_{\mathbf{y}}\left[T \times \max_a \mu_a(\theta_a)\right]$ *is the conventional regret benchmark.*

Empirically (§2.5), we observe that $W^{\textsc{Irs.V-Zero}} \geq W^{\textsc{Irs.V-EMax}}$. In addition, we have $W^{\textsc{Irs.V-EMax}} \geq W^{\mathrm{ideal}}$ since W^{ideal} is the lowest attainable upper bound (Theorem 2.3.1). The second inequality (2.58) holds in a stronger sense: for every outcome ω, the maximal value of the inner problem associated with W^{TS} is greater than that of the inner problem associated with $W^{\textsc{Irs.V-EMax}}$.

While the entire proof is provided in §A.4.3, we highlight here the main ideas. The first result (2.57) follows from the monotonicity of the information structure incorporated in each penalty function: TS, Irs.FH, and Irs.V-Zero replace the realized rewards with $\mathbb{E}(r_t|\boldsymbol{\theta})$, $\mathbb{E}(r_t|\hat{\boldsymbol{\mu}}_{T-1})$, and $\mathbb{E}(r_t|H_{t-1})$, respectively, where $\boldsymbol{\theta}$ is more informative than $\hat{\boldsymbol{\mu}}_{T-1}$, and $\hat{\boldsymbol{\mu}}_{T-1}$ is more informative than H_{t-1} for the DM to infer the value of future reward r_t. Based on this observation, we use a variant of Jensen's inequality to prove the results.[9] The second result (2.58) is proven based on Theorem 4 of [10], which says that if an approximate value function $\widehat{V}$ is a supersolution (Definition A.4.1) to the Bellman equation and a penalty function $\hat{z}$ approximates the ideal penalty with $\widehat{V}$ in place of V^*, the resulting performance bound $W^{\hat{z}}$ is smaller than $\widehat{V}$. By showing that W^{TS} is a supersolution to (2.15), we prove that $W^{\textsc{Irs.V-EMax}} \leq W^{\mathrm{TS}}$ since $z_t^{\textsc{Irs.V-EMax}}$ is constructed upon W^{TS}.

Although Theorem 2.4.1 compares the performance bound among IRS algorithms, we interpret that its tightness, $W^z - V^*$, reflects the degree of optimism that its corresponding policy π^z possesses. Recall that W^z is the expected value of the best possible payoff when the DM is informed of some future outcomes in advance. The weak duality $W^z \geq V^*$ implies that IRS policies are basically optimistic: an IRS policy takes an action as if it can earn more than the optimal policy in the belief that the sampled outcome is the ground truth. In this sense, the gap $W^z - V^*$ captures how optimistically the policy π^z interprets the sampled outcome. When $W^z - V^*$ is relatively small

[9] We remark that $W^{\textsc{Irs.FH}} \geq W^{\textsc{Irs.V-Zero}}$ is not an immediate consequence of the fact that $\sigma(\hat{\boldsymbol{\mu}}_{T-1})$ is a stronger filtration than $\sigma(H_{t-1})$. It further relies on a particular structure of MAB problems: the rewards of an arm are independent and identically distributed conditionally on the parameter. See §A.4.3 for a further discussion.

for a certain penalty function z_t, we may conclude that the penalty function z_t makes the DM less optimistic and induces a policy π^z that performs fewer random explorations.

We further compare the performance of IRS policies using an alternative suboptimality measure. We define the "suboptimality gap" of an IRS policy π^z to be $W^z(T, \mathbf{y}) - V(\pi^z, T, \mathbf{y})$, and analyze it instead of the conventional (Bayesian) regret, $W^{\mathrm{TS}}(T, \mathbf{y}) - V(\pi^z, T, \mathbf{y})$. While its nonnegativity is guaranteed by weak duality (Theorem 2.3.1), more desirably, the optimal policy yields a zero suboptimality gap (Theorem 2.3.1 and Remark 2.3.1). This measure coincides with the conventional regret measure only for TS.

Theorem 2.4.2 (Suboptimality gap for natural exponential families)**.** *Consider an MAB instance with a reward distribution that is a natural exponential family distribution, as described in §2.2.2, in which each arm $a \in \mathcal{A}$ is described with a log-partition function $A_a(\theta_a)$ and a hyperparameter $y_a = (\xi_a, \nu_a)$. Suppose that all the log-partition functions are L-smooth, i.e.,*

$$\frac{d^2}{d\theta_a^2} A_a(\theta_a) \le L, \quad \forall \theta_a \in \Theta_a,\ a \in \mathcal{A}. \tag{2.59}$$

Further assume that $\nu_a = \nu$ for all $a \in \mathcal{A}$. Then, for any $T \ge 2$, we have

$$W^{\mathrm{TS}}(T, \mathbf{y}) - V(\pi^{\mathrm{TS}}, T, \mathbf{y}) \le 2\sqrt{L}\left[\frac{1}{\sqrt{\nu}} + \sqrt{2\log T} \times \left(\frac{K}{\sqrt{\nu}} + 2\sqrt{KT}\right)\right], \tag{2.60}$$

$$W^{\textsc{Irs.FH}}(T, \mathbf{y}) - V(\pi^{\textsc{Irs.FH}}, T, \mathbf{y}) \le 2\sqrt{L}\left[\frac{1}{\sqrt{\nu}} + \sqrt{2\log T} \times \left(\frac{K}{\sqrt{\nu}} + 2\sqrt{KT} - \frac{1}{3}\sqrt{\frac{T}{K}}\right)\right], \tag{2.61}$$

$$W^{\textsc{Irs.V-Zero}}(T, \mathbf{y}) - V(\pi^{\textsc{Irs.V-Zero}}, T, \mathbf{y}) \le \sqrt{L}\left[\frac{1}{\sqrt{\nu}} + \sqrt{2\log T} \times \left(\frac{K}{\sqrt{\nu}} + 2\sqrt{KT} - \frac{1}{3}\sqrt{\frac{T}{K}}\right)\right]. \tag{2.62}$$

Remark 2.4.2. *For a Bernoulli MAB with symmetric arms, each of which has a prior Beta(α, β) for its mean reward, we have $L = \frac{1}{2}$ and $\sqrt{\nu} = \sqrt{\alpha + \beta}$.*

Remark 2.4.3. *For a Gaussian MAB with symmetric arms, each of which has a prior $\mathcal{N}(m, v^2)$*

for its mean reward and a noise variance σ^2, we have $L = \sigma$ and $\sqrt{v} = v/\sigma$.

Theorem 2.4.2 indirectly shows to the improvements to the suboptimality gaps: although all the bounds have the same asymptotic order of $O(\sqrt{KT \log T})$, the IRS policies improve the leading coefficient or the additional term.[10] These results hold for a wide range of MAB problems including the Bernoulli MAB and the Gaussian MAB as stated in Remarks 2.4.2 and 2.4.3.

The proof of Theorem 2.4.2, provided in §A.4.4, relies on an essential property of IRS policies that generalizes the "probability matching" property of TS, i.e., a matching between nature's randomness and the decision maker's randomness. It is well known that TS is randomized in a way that, conditional on past observations, the probability that an action a is chosen equals the probability that the action a is chosen by someone who knows the parameters. Analogously, the IRS policy π^z is randomized in a way that, conditional on past observations, the probability that an action a is chosen equals the probability that the action a is chosen by someone who knows the entire future but is penalized (Proposition A.4.4). Recall that the penalties are designed to penalize the benefit from having additional future information. A better choice of penalty function would prevent the policy π^z from picking an action that is overly optimized for a randomly sampled future realization, which in turn would improve the quality of the decision making.

Given the above observation, our proof utilizes the approach taken by [21] that exploits the probability matching property of TS to bound its Bayesian regret. More specifically, for each penalty function, we carefully construct a sequence of confidence intervals on the mean reward such that the corresponding policy's instantaneous suboptimality at each time (loss against the

[10] Recall that $W^{\mathrm{TS}} - V(\pi^{\mathrm{TS}})$ represents the Bayesian regret of TS. It will be worth mentioning some known results that may be comparable to the bound (2.60) established in Theorem 2.4.2.

For the cases where the reward distributions have a bounded support in $[0, 1]$, [19] have shown that the Bayesian regret of TS is bounded from above by $14\sqrt{KT}$; and further shown that its asymptotic order is unimprovable in the sense that for any policy there exists a prior distribution such that the policy experiences Bayesian regret no smaller than $\frac{1}{20}\sqrt{KT}$. However, this does not imply that the regret of the Bayesian optimal policy is bounded from below by $\frac{1}{20}\sqrt{KT}$ in the context of Theorem 2.4.2, since we consider a specific prior and the policy optimized to that prior.

For Gaussian MAB in the non-Bayesian setting, [18] have shown that the regret of TS is $O(\sqrt{KT \log T})$; and further shown that its asymptotic order is unimprovable in the sense that for any policy there exists an instance (i.e., the set of true mean values) such that the policy's regret is at least $\Omega(\sqrt{KT \log K})$.

While there is no result in the literature that is comparable to the other bounds (2.61) and (2.62), we conjecture that they will be tight just as the bound for TS (2.60) is, given the fact all three policies exhibit the identical asymptotic behavior for large T (Proposition 2.4.1).

hindsight solution) is bounded by the width of the confidence interval approximately. For a better penalty function, the confidence intervals can be made tighter so that the total suboptimality can also be bounded more effectively. In our analysis, the natural exponential family is assumed in order to analyze the concentration of posterior distribution in a closed form, and the smoothness condition on the log-partition function is assumed in order to guarantee that the reward distribution is sub-Gaussian, whereas [21] consider an arbitrary reward distribution with a bounded support.

2.5 Numerical experiments

2.5.1 Experimental setup

We conduct numerical simulations to evaluate the effectiveness of our framework in comparison to alternative algorithms. In addition to the IRS algorithms discussed so far, we consider other recently developed algorithms that are particularly suitable for a Bayesian setting: the Bayesian upper confidence bound [23] (BAYES-UCB, with a quantile of $1 - \frac{1}{t}$), information-directed sampling [22] (IDS), the optimistic Gittins index [26] (OGI, one-step look-ahead approximation with a discount factor of $\gamma_t = 1 - \frac{1}{t}$), and the Lagrangian index policies suggested in [27] (LAGR-RT and LAGR-OT, with a random and an optimal tie-breaking rule, respectively).

Our numerical experiments include Beta-Bernoulli MABs and Gaussian MABs. Given an MAB problem instance specified by the prior distribution $\mathcal{P}(\mathbf{y})$ and the reward distribution $\mathcal{R}$, we simulate the policies and calculate the IRS bounds with respect to the different values of time horizon T.

Let S be the number of simulations we perform. For each $s \in \{1, \dots, S\}$, we first sample the parameters $\theta_a^{(s)} \sim \mathcal{P}_a(y_a)$ and the rewards $R_{a,n}^{(s)} \sim \mathcal{R}_a(\theta_a^{(s)})$ for all $n \in \{1, \dots, T_{\max}\}$ and $a \in \mathcal{A}$, which is equivalent to sampling an outcome $\omega^{(s)} \sim \mathcal{I}(\mathbf{y})$. Given the s^{th} sampled outcome $\omega^{(s)}$, for each time horizon $T \in \{5, 10, 15, \dots, T_{\max}\}$, we simulate each policy π (that may utilize the time horizon T); i.e., at each time $t = 1, \dots, T$, the policy makes a decision[11] on which arm to pull,

[11] Recall that IRS policies are randomized policies that perform their own simulations at each time along the sample path. This posterior sampling procedure is independent of the random generation of true outcomes.

A_t^π, and then the associated reward, $r_t(\mathbf{A}_{1:t}^\pi, \omega^{(s)}) = R^{(s)}_{A_t^\pi, n_t(\mathbf{A}_{1:t}^\pi, A_t^\pi)}$, is revealed accordingly. After simulating one sample path, $\sum_{t=1}^T \mu_{A_t^\pi}(\theta^{(s)}_{A_t^\pi})$ is recorded as the performance of π for the s^{th} sample, and the expected performance $V(\pi, T, \mathbf{y})$ is measured by its sample average across S samples for each T.

In order to compute IRS bounds, we use the same set of samples $\omega^{(1)}, \ldots, \omega^{(S)}$. For each penalty function z and for each $T \in \{5, 10, \ldots, T_{\text{max}}\}$, we solve the associated inner problems with respect to $\omega^{(1)}, \ldots, \omega^{(S)}$, and the IRS bound $W^z(T, \mathbf{y})$ is evaluated by taking the average of the maximal values over S instances.

More explicitly, we use the following sample averages to calculate $V(\pi, T, \mathbf{y})$ and $W^z(T, \mathbf{y})$:

$$V(\pi, T, \mathbf{y}) \approx \frac{1}{S} \sum_{s=1}^S \left(\sum_{t=1}^T \mu_{A_t^\pi}(\theta^{(s)}_{A_t^\pi}) \right), \quad W^z(T, \mathbf{y}) \approx \frac{1}{S} \sum_{s=1}^S \max_{\mathbf{a}_{1:T} \in \mathcal{A}^T} \left\{ \sum_{t=1}^T r_t(\mathbf{a}_{1:t}, \omega^{(s)}) - z_t(\mathbf{a}_{1:t}, \omega^{(s)}) \right\}. \tag{2.63}$$

Note again that the same outcome $\omega^{(s)}$ is used across the different values of time horizon T and across different algorithms. Sharing the randomness enhances the consistency of the estimates. In what follows, we use 20,000 samples (i.e., $S = 20,000$).

Based on $V(\pi, T, \mathbf{y})$ and $W^{\text{TS}}(T, \mathbf{y})$ measured with the sample averages, we calculate the Bayesian regret of a policy π:

$$\text{BayesRegret}(\pi, T, \mathbf{y}) \triangleq \mathbb{E}\left[\sum_{t=1}^T \max_a \mu_a(\theta_a) - \mu_{A_t^\pi}(\theta_{A_t^\pi}) \right] = W^{\text{TS}}(T, \mathbf{y}) - V(\pi, T, \mathbf{y}), \tag{2.64}$$

which is a conventional measure in performance analysis of Bayesian algorithms as discussed in §2.3.1. We further calculate the regret (lower) bound obtained from a IRS penalty function z_t:

$$\text{RegretBound}(z, T, \mathbf{y}) \triangleq W^{\text{TS}}(T, \mathbf{y}) - W^z(T, \mathbf{y}). \tag{2.65}$$

By the weak duality (Theorem 2.3.1), we have $\text{BayesRegret}(\pi, T, \mathbf{y}) \geq \text{RegretBound}(z, T, \mathbf{y})$ for any $\pi \in \Pi_{\mathbb{F}}$. By its definition, the regret bound produced by TS is zero.

2.5.2 Results

Bernoulli MAB with two arms ($K = 2$). We first provide the results for a Bernoulli MAB in which

$$\mu_a \sim \text{Beta}(1,1), \quad R_{a,n} \sim \text{Bernoulli}(\mu_a), \quad \forall a \in \{1,2\}. \tag{2.66}$$

We consider relatively short time horizons ($\leq T_{\max} = 200$) since we are focusing on a finite-horizon regime rather than an asymptotic regime. In this particular case, since the state (belief) space is discrete and small in size, $O(T^4)$, we are able to solve the Bellman equations (2.15) numerically, and thus we can implement the Bayesian optimal policy, which is labeled as OPT in what follows.

Figure 2.2 shows the regrets (solid lines) of all the policies discussed above and the regret bounds (dashed lines) produced by the IRS algorithms.[12] Table 2.3 provides further details including the percentage improvement in regret over TS, i.e.,

$$\text{RegretImprovement}(\pi, T, \mathbf{y}) \triangleq 1 - \frac{\text{BayesRegret}(\pi, T, \mathbf{y})}{\text{BayesRegret}(\text{TS}, T, \mathbf{y})},$$

and the improvement in regret bound over TS benchmarked to the regret of the best performing algorithm, i.e.,

$$\text{BoundImprovement}(\pi, T, \mathbf{y}) \triangleq \frac{\text{RegretBound}(z, T, \mathbf{y}) - \text{RegretBound}(z^{\text{TS}}, T, \mathbf{y})}{\min_{\pi'} \text{BayesRegret}(\pi', T, \mathbf{y})}.$$

In Figure 2.2, note that lower regret curves are better, and higher bound curves are better.

Comparing the IRS algorithms (TS, IRS.FH, IRS.V-ZERO, IRS.V-EMAX, and OPT), we first observe a clear improvement in both the performance of policies and the tightness of bounds, as we adopt a more complicated penalty function, albeit one that requires a longer run time: as visualized in Figure 2.2, the regret curve approaches the OPT curve from above and the bound curve approaches it from below, where the OPT curve represents the lowest attainable regret that

[12] There also exists a performance bound induced by the Lagrangian index policies. We omit it from Figure 2.2, however, since that bound is not so tight and thus not informative to be displayed in the same plot; e.g., when $T = 200$, the associated regret bound is -12.54, which is far below the current x-axis.

is the highest attainable regret bound at the same time. The suboptimality gap (the gap between a regret curve and its corresponding bound curve) becomes smaller, which is consistent with the implication of Theorem 2.4.2.

Finally, we note that the IRS.INDEX policy is outperforming all the other policies; i.e., the regret curve of IRS.INDEX is surprisingly close to the OPT curve. Although it is developed based on IRS.V-EMAX, it performs better than IRS.V-EMAX, and the reasons for that still need to be researched.

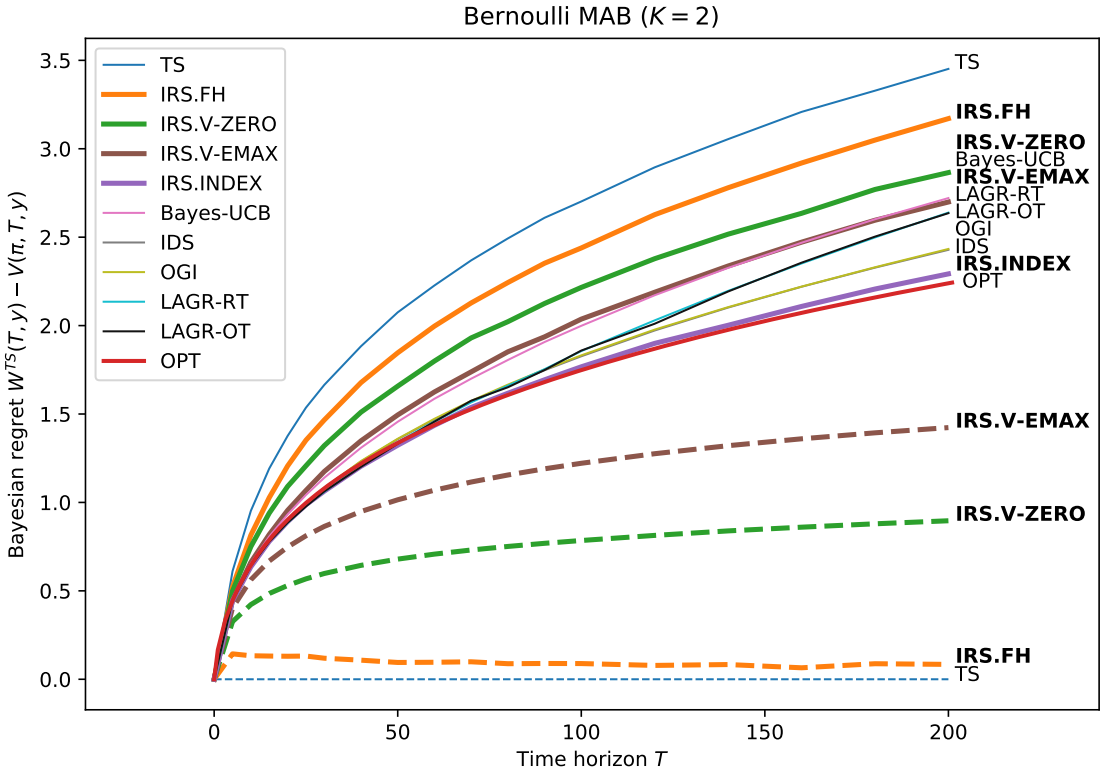

Figure 2.2: Regret plot for a Bernoulli MAB with two arms. The solid lines represent the (Bayesian) regret of algorithms, $W^{\mathrm{TS}}(T, \mathbf{y}) - V(\pi, T, \mathbf{y})$, and the dashed lines represent the regret bounds that IRS algorithms produce, $W^{\mathrm{TS}}(T, \mathbf{y}) - W^{z}(T, \mathbf{y})$. Each data point reports the average across 20,000 simulations.

Algorithm	Bayesian regret (s.e.)	Regret improvement	Regret lower bound (s.e.)	Bound improvement	Policy run time
TS	3.45 (0.021)	0.0%	0.00 (–)	0.0%	17 ms
IRS.FH	3.17 (0.020)	8.1%	0.08 (0.040)	3.8%	37 ms
IRS.V-ZERO	2.87 (0.021)	17.0%	0.90 (0.055)	40.0%	527 ms
IRS.V-EMAX	2.70 (0.020)	21.8%	1.42 (0.326)	63.6%	29.5 sec
IRS.INDEX	2.29 (0.023)	33.6%	–	–	3.6 sec
BAYES-UCB	2.72 (0.020)	21.2%	–	–	44 ms
IDS	2.43 (0.028)	29.6%	–	–	3.7 sec
OGI	2.43 (0.028)	29.5%	–	–	262 ms
LAGR-RT	2.64 (0.046)	23.5%	-12.54	–	19 ms*
LAGR-OT	2.64 (0.046)	23.6%	-12.54	–	14 ms*
OPT	**2.24** (–)	35.1%	**2.24** (–)	100.0%	–

Table 2.3: Simulation results for a Bernoulli MAB with two arms when $T = 200$. The best results are emphasized with bold letters. The third and fifth columns show the percentage improvements over TS in regret and in bound respectively; e.g., IRS.V-EMAX achieves a regret that is 21.8% better than that of TS, and yields a regret bound that accounts for 63.7% of the lowest regret observed empirically.
The last column shows the average time required for a policy to make decisions along one sample path including the time required posterior sampling for the case of IRS policies. *LAGR-RT and LAGR-OT require substantial offline computation prior to simulation. This takes around 20 hours in the setting of this simulation.

Bernoulli MAB with ten arms ($K = 10$). We next consider a Bernoulli MAB with ten arms and $T_{\max} = 500$. IRS.V-EMAX and OPT are omitted from this simulation due to their computational cost, and so are LAGR-RT and LAGR-OT for long horizons[13] ($T > 350$). Figure 2.3 and Table 2.4 show the simulation results. We again observe a monotonic improvement in the performance of policies and the tightness of bounds among IRS algorithms, and the IRS.INDEX policy still

[13] LAGR-RT and LAGR-OT require substantial offline pre-computation. This involves a convex optimization problem with T decision variables, where a single evaluation of the objective function requires $\Theta(T^3)$ operations. As recommended by [27], we have implemented a cutting-plane method using a commercial optimization software (Gurobi), but it takes over a week to complete the pre-computation when $T = 350$.

performs best.

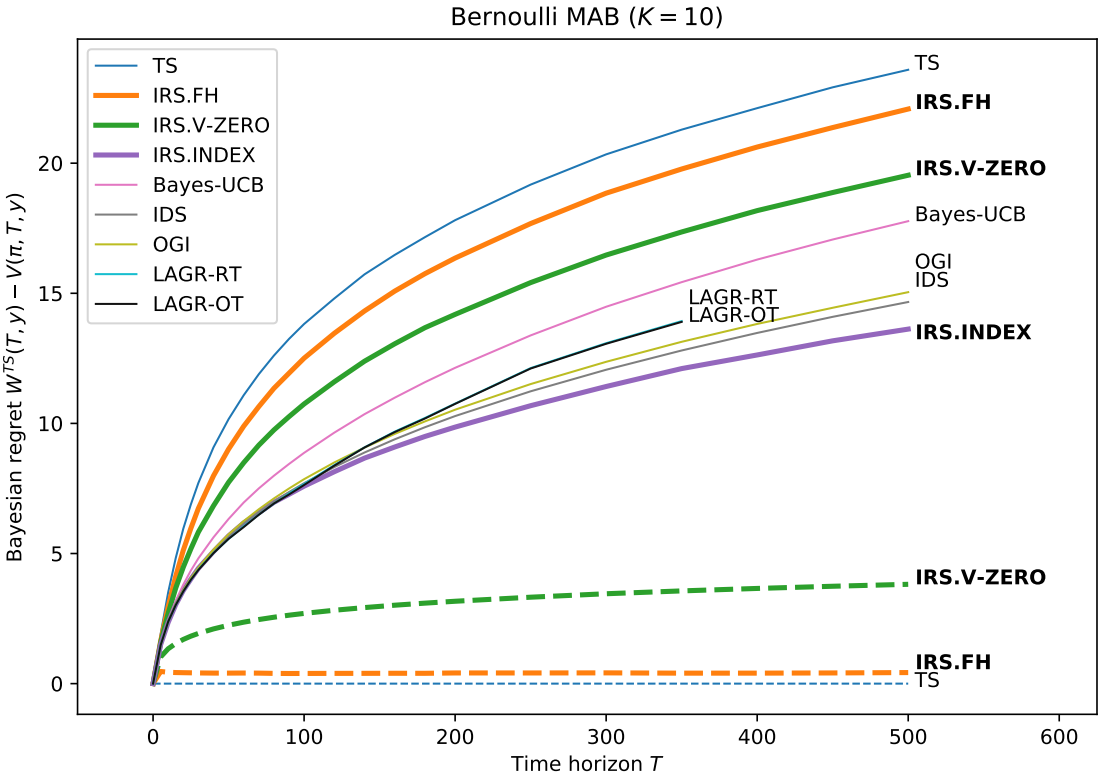

Figure 2.3: Regret plot for a Bernoulli MAB with ten arms. LAGR-RT and LAGR-OT are simulated only for $T \leq 350$ due to the computational cost (see Footnote 13).

Algorithm	Bayesian regret (s.e.)	Regret improvement	Regret lower bound (s.e.)	Bound improvement	Policy run time
TS	23.59 (0.078)	0.0%	0.00 (–)	0.0%	50 ms
IRS.FH	22.08 (0.076)	6.4%	0.43 (0.042)	4.0%	300 ms
IRS.V-ZERO	19.54 (0.074)	17.2%	**3.82** (0.058)	35.6%	17.0 sec
IRS.INDEX	**13.62** (0.080)	42.2%	–	–	56.2 sec
BAYES-UCB	17.77 (0.077)	24.7%	–	–	140 ms
IDS	14.67 (0.093)	37.8%	–	–	16.4 sec
OGI	15.04 (0.092)	36.2%	–	–	2.6 sec

Table 2.4: Simulation results for a Bernoulli MAB with ten arms when $T = 500$.

Gaussian MABs ($K = 2$ or 10). We next consider Gaussian MABs in which

$$\mu_a \sim \mathcal{N}(0, 1^2), \quad R_{a,n} \sim \mathcal{N}(\mu_a, 1^2), \quad \forall a \in \{1, \ldots, K\}. \tag{2.67}$$

Figure 2.4 and Table 2.5 show the case of two arms ($K = 2$), and Figure 2.5 and Table 2.6 show the case of ten arms ($K = 10$). The algorithms LAGR-RT and LAGR-OT are not implemented for Gaussian MABs since they require either discrete belief states or some form of state discretization. The results are similar to those of Bernoulli MABs.

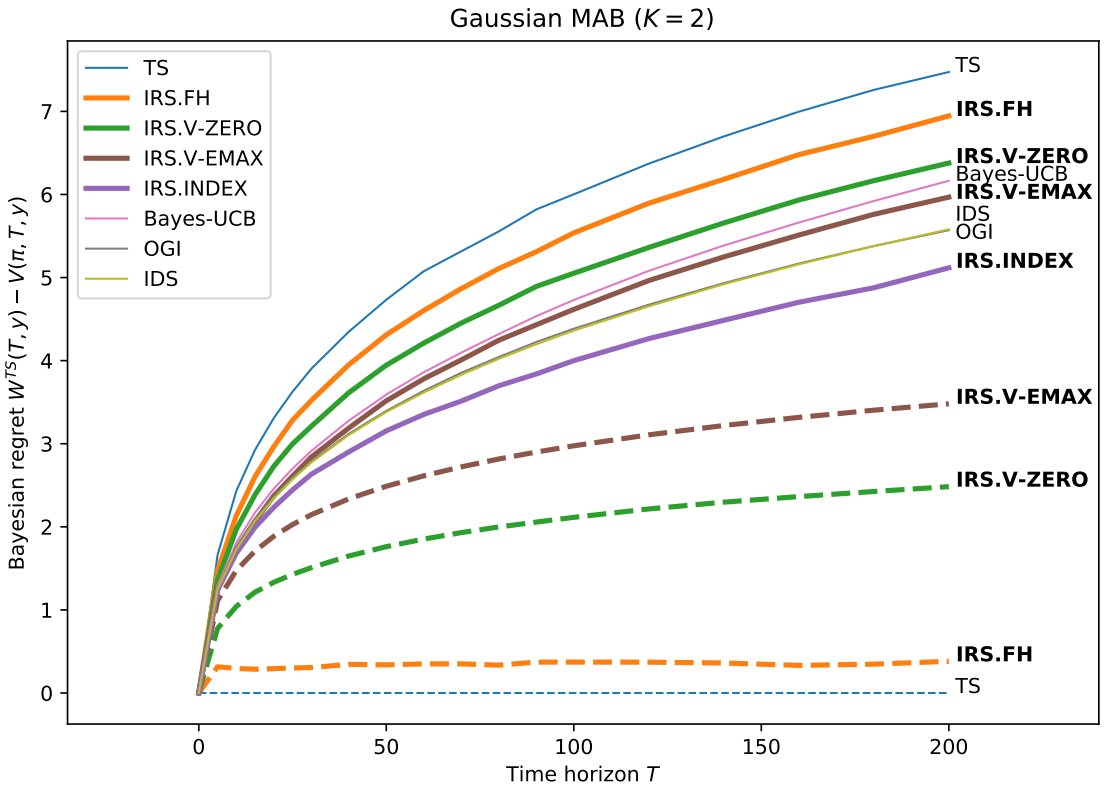

Figure 2.4: Regret plot for a Gaussian MAB with two arms.

Algorithm	Bayesian regret (s.e.)	Regret improvement	Regret lower bound (s.e.)	Bound improvement	Policy run time
TS	7.47 (0.047)	0.0%	0.00 (–)	0.0%	17 ms
IRS.FH	6.94 (0.045)	7.1%	0.38 (0.100)	7.4%	37 ms
IRS.V-ZERO	6.38 (0.048)	14.7%	2.48 (0.133)	48.5%	625 ms
IRS.V-EMAX	5.97 (0.044)	20.2%	**3.48** (1.154)	68.0%	13.3 sec
IRS.INDEX	**5.12** (0.054)	31.5%	–	–	2.2 sec
BAYES-UCB	6.16 (0.045)	17.5%	–	–	38 ms
IDS	5.58 (0.068)	25.3%	–	–	679 ms
OGI	5.57 (0.067)	25.5%	–	–	196 ms

Table 2.5: Simulation results for a Gaussian MAB with two arms when $T = 200$.

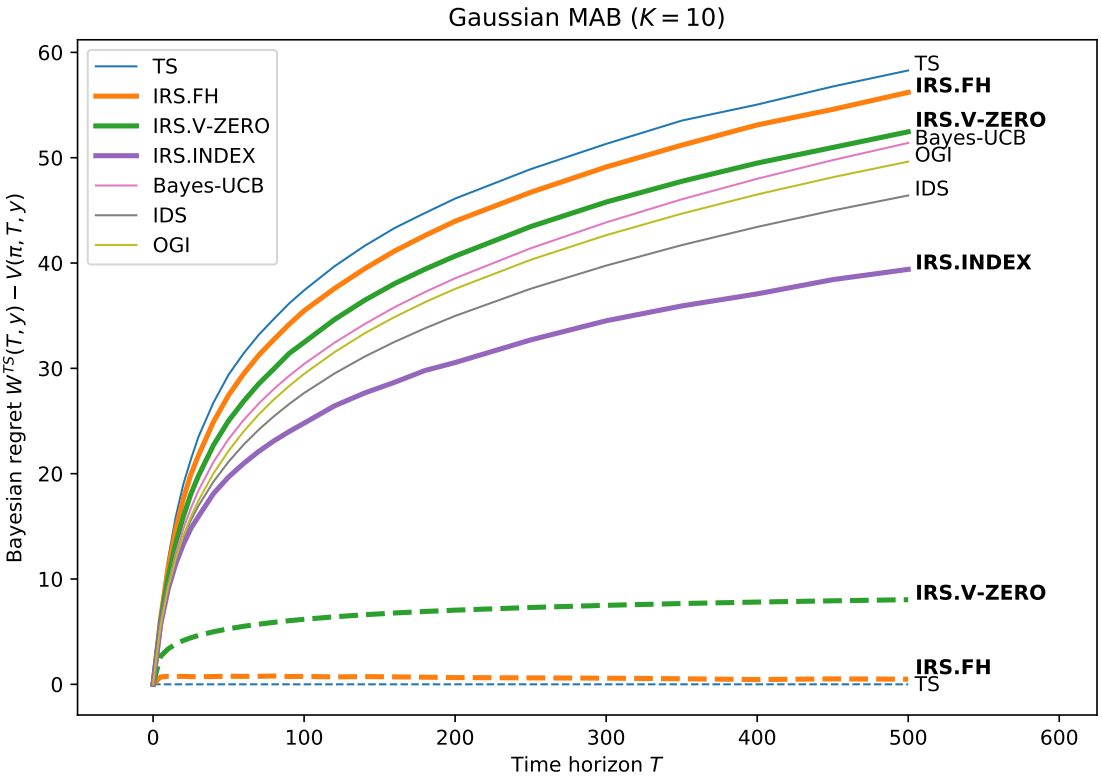

Figure 2.5: Regret plot for a Gaussian MAB with ten arms.

Algorithm	Bayesian regret (s.e.)	Regret improvement	Regret lower bound (s.e.)	Bound improvement	Policy run time
TS	58.28 (0.180)	0.0%	0.00 (–)	0.0%	35 ms
IRS.FH	56.20 (0.180)	3.6%	0.48 (0.156)	1.2%	215 ms
IRS.V-ZERO	52.46 (0.188)	10.0%	**8.04** (0.216)	20.4%	13.7 sec
IRS.INDEX	**39.40** (0.244)	32.4%	–	–	30.4 sec
BAYES-UCB	51.40 (0.178)	11.8%	–	–	77 ms
IDS	46.41 (0.324)	20.4%	–	–	4.0 sec
OGI	49.63 (0.335)	14.8%	–	–	1.6 sec

Table 2.6: Simultion results for a Gaussian MAB with ten arms when $T = 500$.

Gaussian MAB with different noise variances ($K = 5$). We next consider a problem where

$$\mu_a \sim \mathcal{N}(0, 1^2), \quad R_{a,n} \sim \mathcal{N}(\mu_a, \sigma_a^2), \quad \forall a \in \{1, \dots, 5\} \tag{2.68}$$

and $(\sigma_1, \sigma_2, \sigma_3, \sigma_4, \sigma_5) = (0.1,\ 0.4,\ 1,\ 4,\ 10)$. In this MAB instance, it is particularly crucial for the algorithms to consider how much the DM can learn about each of the arms during the remaining time periods, since the difficulty of estimating the mean reward of an arm a heavily depends on the noise level σ_a that varies across the arms.[14]

As shown in Figure 2.6, BAYES-UCB shows a particularly poor performance, as it keeps pulling arm 5 without considering the fact that arm 5 is too noisy to be learnt within such a short period of time (i.e., $T \leq 500$). By contrast, we observe that our IRS policies and IDS algorithm outperform the BAYES-UCB, OGI, and TS algorithms, since they explicitly take into account the value of exploration by quantifying the informativeness of a new observation for each arm (more specifically, by considering how the belief will change as a new reward realization is revealed). Notably, the IRS.FH policy, which is a very simple modification of TS, significantly improves the performance of TS without degrading its computational efficiency.

The example also illustrates the significance of having a tighter performance bound. If the benchmark is set to $W^{\text{IRS.V-ZERO}}$, when $T = 500$, the IRS.INDEX* policy[15] achieves 94% $\left(= \frac{V(\pi^{\text{IRS.INDEX*}}, T, \mathbf{y})}{W^{\text{IRS.V-ZERO}}(T, \mathbf{y})}\right)$

[14] In order for the posterior distribution to be concentrated so as to have a standard deviation of 0.1, for example, one observation is enough for arm 1 whereas 100 and 10,000 observations are required for arm 3 and arm 5, respectively.

[15] The IRS.INDEX* policy is a heuristic modification of the IRS.INDEX policy. See §A.2.3.

of the benchmark. If the benchmark is set to W^{TS} instead, as in a conventional regret analysis, we might have concluded that the IRS.INDEX* policy achieves only 88% $\left(= \frac{V(\pi^{\text{IRS.INDEX*}}, T, \mathbf{y})}{W^{\mathrm{TS}}(T, \mathbf{y})}\right)$ of that (looser) bound, which would suggest a larger margin of possible improvement.

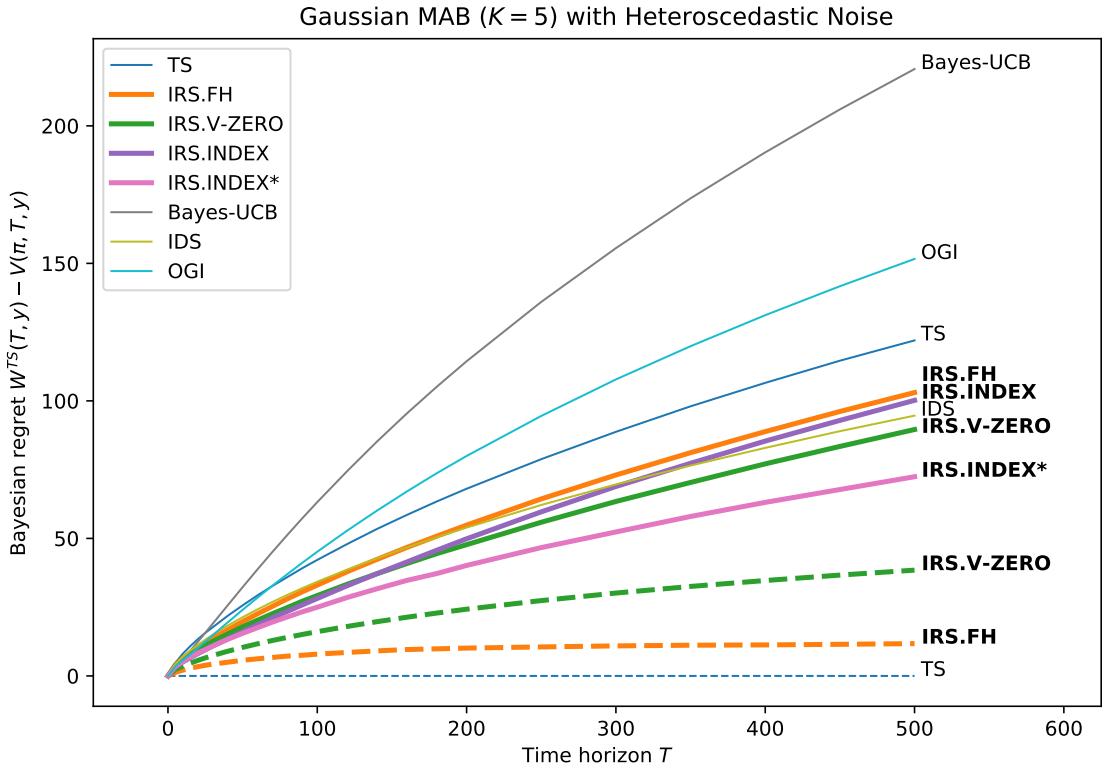

Figure 2.6: Regret plot for a Gaussian MAB with five arms with different noise variances.

Algorithm	Bayesian regret (s.e.)	Regret improvement	Regret lower bound (s.e.)	Bound improvement	Policy run time
TS	121.99 (0.615)	0.0%	0.00 (–)	0.0%	34 ms
IRS.FH	103.03 (0.628)	15.5%	11.75 (0.656)	16.2%	128 ms
IRS.V-ZERO	89.59 (0.690)	26.6%	**38.47** (0.827)	53.1%	7.4 sec
IRS.INDEX	100.20 (0.657)	17.9%	–	–	12.8 sec
IRS.INDEX*	**72.43** (0.866)	40.6%	–	–	12.3 sec
BAYES-UCB	220.66 (1.285)	-80.9%	–	–	88 ms
IDS	94.63 (0.817)	22.4%	–	–	2.9 sec
OGI	151.61 (1.030)	-24.3%	–	–	829 ms

Table 2.7: Simulation results for a Gaussian MAB with five arms with different noise variances when $T = 500$.

2.6 Extensions

Below, we describe several natural generalizations of the methods developed in this paper beyond the setting of Section 2.2:

MAB with unknown time horizon. This paper studies finite-time horizon MABs for which we suggest algorithms that exploit the knowledge of the time horizon T and we focus on a relatively small T such that the time horizon becomes an important ingredient in optimally balancing exploration and exploitation. We briefly illustrate how to relax our framework's dependency on T, i.e., how to extend to the setting with an unknown horizon and the setting with an indefinitely long horizon.

First, our framework (penalties, policies, and upper bounds) can naturally incorporate the unknown T within the *Bayesian setting*; i.e., the horizon T is also a random variable whose prior distribution is known. As a simple case, if T is independent of the DM's actions, we can reformulate the objective function of the inner problem as $\sum_{t=1}^{\infty} \gamma_t \left(r_t(\mathbf{a}_{1:t}, \omega) - z_t(\mathbf{a}_{1:t}, \omega) \right)$ where the discount factor $\gamma_t \triangleq \mathbb{P}[T \geq t]$ is the survivor probability, and $r_t(\cdot)$ and $z_t(\cdot)$ are the reward and

penalty terms used in the paper. Alternatively, we can treat the random variable T like the random reward realizations by sampling T from its prior distribution while a penalty function (additionally) penalizes for the gain from knowing T (one can imagine that the outcome ω now includes the realization of T and not only the future reward realizations). Structural results such as weak duality and strong duality will continue to hold.

Second, we can consider a practical modification of IRS policies when T is large or infinite. We can construct a dual feasible penalty function that mixes IRS.FH and IRS.V-ZERO,[16] which induces an algorithm whose complexity is $O\left(K \min\{T, T_0\}^2\right)$ for some predefined constant T_0. Alternatively, we can convert the IRS.V-EMAX or IRS.INDEX policy into an anytime policy by setting the inner problem's horizon large enough, despite that the performance bound will no longer be obtainable.

MAB in more complicated settings. Even though this paper develops a framework for the stochastic MAB with independent arms, which would be the simplest and oldest problem in the MAB literature, we believe that our framework applies to more complicated settings. Consider the following examples:

- A finite-horizon MAB with correlated arms (e.g., $R_{a,n} \sim \mathcal{N}(\mathbf{x}_a^\top \boldsymbol{\theta}, \sigma_a^2)$ where $\boldsymbol{\theta} \in \mathbb{R}^d$ is shared across the arms, and $\mathbf{x}_a \in \mathbb{R}^d$ is an arm's feature vector): IRS.V-ZERO can be immediately implemented by adopting the DP algorithm discussed in §A.2.2.

- MAB with the delayed reward realization: IRS.FH can be immediately implemented by simulating the DM's learning process in the presence of delay.

- MAB with a budget constraint (in which each arm consumes a certain amount of budget and the DM wants to maximize the total reward within a limited budget. See [28]): all IRS algorithms can be implemented by solving a budget-constrained optimization problem instead of a horizon-constrained optimization problem.

[16] In its inner problem, IRS.V-ZERO-like penalties are applied for the initial $\lfloor T_0/K \rfloor$ pulls and then IRS.FH-like penalties are applied for the later pulls.

In these extensions, we obtain not only the online decision-making policies but also their performance bounds as in this paper. Generally speaking, our framework provides a systematic way of improving TS by taking into account the exploitation-exploration trade-off more carefully, particularly in the presence of some constraint that induces incomplete learning; the main challenge would be to design a suitable penalty function that is tractable yet captures the problem-specific exploration-exploitation trade-off precisely.

2.7 Conclusion

Contribution to MAB literature. We first highlight that our IRS framework generalizes Thompson sampling to the finite-horizon MAB setting. As pointed out in [29], TS may perform poorly in time-sensitive learning problems in which exploitation is rather more encouraged than exploration. Interpreted as a special case of IRS policies, it is clear that TS is implicitly assuming an infinite time horizon in the sense that its associated inner problem solves a best-arm identification problem with an infinite number of observations. As summarized in Table 2.2, IRS algorithms consider more complicated inner problems in which the benefit from exploration is limited by the time-horizon constraint. While maintaining the Bayesian recursive structure of its sequential decision-making process, we improve TS within a unified framework that also includes the Bayesian optimal policy as another special case.

Furthermore, the IRS framework provides a set of (Bayesian) performance bounds that are tighter than the conventional benchmark that has been widely used since [16]. We believe that these benchmarks would be useful, in a Bayesian setting, in measuring the optimality of an algorithm or in assessing the intrinsic difficulty of an MAB problem instance.

Contribution to information relaxation literature. The information relaxation framework is certainly a powerful tool to obtain performance bounds in a general class of decision-making problems. Although there have been several studies [9] that elicit a decision-making policy based on this framework, they are limited to using a performance bound as a proxy for the value function. Instead of approximating the value function explicitly, the IRS framework considers simulation-

based randomized policies that make each decision that is optimized to a single instance of a simulated environment, and our results show that this scheme is very powerful in online learning problems where random exploration is required.

In applying the information relaxation framework to a particular application, the most challenging task is to find a suitable penalty function that is tractable yet yields a tight performance bound. In this paper, by exploiting the recursive structures embedded in the Bayesian learning process, we derive a series of penalty functions so that users themselves can find a balance between the quality of policies/bounds and the computational cost. We also provide theoretical analyses of the tightness of performance bounds and the suboptimality of associated policies by leveraging the existing analysis developed in the MAB literature. These analytic results would be rare in the information relaxation literature due to the complex nature of the performance bound produced by the information relaxation framework.

Chapter 3: Policy Gradient Optimization of Thompson Sampling Policies

3.1 Introduction

In both academia and industry, Thompson sampling has emerged as a leading approach to exploration in online decision making. This is driven by the algorithm's simplicity, generality, ability to leverage rich prior information about problem, and its resilience to delayed feedback. But, like most popular bandit algorithms, it is a heuristic design based on intuitive appeal and some degree of mathematical insight. The tutorial paper by [30] details numerous settings in which Thompson sampling can be grossly suboptimal. We highlight several such situations:

- Settings where the time horizon is short relative to the number of arms. As an extreme case, in the situation there is a single period remaining, the myopic policy is optimal and Thompson sampling will over explore. At another extreme, if there are many arms, it may be optimal to only restrict exploration to a subset so that a good arm can be identified in time to exploit over a reasonable time frame.

- Thompson sampling does not directly consider reward noise. If there is a significant heterogeneity in the noise across arms, Thompson sampling may suboptimally pull noisy arms about which there is little hope to learn.

- In settings with correlated arms, pulling a single arm may provide information about many other arms. In these settings, there may be "free exploration" where, for example, a myopic policy might learn about all arms and the type of explicit exploration undertaken by Thompson sampling may be wasteful.

An underlying theme in the above example is the fact that Thompson sampling does not make an explicit exploration-exploitation trade off. Even in less extreme settings, Thompson sampling is

generally thought to explore too aggressively.

Thompson sampling is designed for a Bayesian multi-armed bandit problem, a well-defined optimization problem that has long been approached using the tools of dynamic programming. Despite this, the literature offers no way to use computation, rather than human ingenuity, to improve on standard Thompson sampling. We propose and benchmark the use of policy gradient methods to optimize over a given family of Thompson sampling style algorithms. The proposed methods use substantial offline computation, but the resulting policies can be executed without additional real-time computation.

At first glance, it appears standard policy gradient algorithms [31] cannot be efficiently applied to Thompson sampling. The challenge is that traditional policy gradient methods require computation of the score function of the distribution of actions. While, in principle, Thompson sampling randomly draws an action at each decision point, the distribution over actions from which it samples is not available in closed form. Instead, efficient implementations sample a model parameter from a posterior distribution and then select the action that is optimal under this sampled parameter. Under such an implementation, it may be difficult to compute the probability of selecting each action.

Our central insight is as follows: we view the posterior parameter sampled by Thompson sampling as a kind of "pseudo-action". In our framework, a *sampling policy* maps the history of observations to a probability distribution over the parameter space. A full algorithm will, in each period, draw a sample from according to the sampling policy and subsequently apply the base action that is optimal under the sampled parameter. In standard Thompson sampling, the sampling policy applies Bayes' rule, mapping any history to the associated posterior distribution over pseudo-actions (parameters). Mathematically, this is equivalent to the standard formulation which views the decision as a choice of base action. Critically, however, by viewing the decision as a choice of pseudo-action (parameter), the distribution of pseudo-actions for Thompson sampling is often available in closed-form: it is simply the posterior distribution of the parameter. This simple but powerful shift in perspective enables us to search over Thompson sampling style policies using

policy gradient methods. Indeed, we will use policy gradient to search over a class of sampling policies that are themselves parameterized by hyper-parameters we call *meta-parameters*.

A sampling policy could be parameterized in many ways. One option is to parameterize them by complex neural networks. Our experiments demonstrate that even simple modifications of standard Thompson sampling offer substantial benefit. One approach builds on Thompson sampling by viewing the statistical parameters of the Bayesian model (e.g., the prior distribution, the noise distribution, etc.) as meta-parameters. Another takes the posterior distribution used by standard Thompson sampling and reshapes it. Policy gradient methods for searching over the meta-parameters are tractable as long as (i) the sampling policies can be applied efficiently and (ii) given any history, one can efficiently calculate derivatives of the sampling distribution's log density with the respect to the meta-parameters. Typically (ii) requires that probability density function is known up to a proportionality constant.

Our work has a close conceptual connection to work on meta-learning [see e.g., 32, 33]. As is nicely articulated by [34], many companies face a large sequence of experimentation tasks, raising the question of how to effectively share information across these tasks. Consider a web company who may run thousands of A/B tests per year, giving them strong prior knowledge of the distribution of effect sizes and click through rates. Or a news article recommendation service has a new set of articles each day and needs to experiment to learn which will be popular. Each day can be viewed as its own instance of a bandit problem and the platform's goal is to do well on average across a large number of days. [34] suggest an empirical Bayesian approach, where the prior of Thompson sampling is statistically estimated from data on previous tasks. This view of meta-learning as learning a prior distribution has long been recognized. Our approach, however, will not be to apply Thompson sampling directly using some form of statistically learned prior, since Thompson sampling is not itself an optimal policy. If historical data can be used to build a simulator of this meta-bandit problem, then it is more appropriate to search over Thompson sampling like policies aiming to directly optimize the true performance metric, the average reward. This idea — shifting from learning elements of the statistical model such as the prior distribution

or noise model via statistical estimation to direct optimization — may especially be powerful in settings where the statistical model is mis-specified.

Our contributions are as follows:

1. **We develop a tractable framework for policy gradient estimation for sampling policies.**

   Several recent works have explored the use of gradient based search to tune bandit algorithms [35, 36]. Relative to these works, one of our main contributions is to uncover a way to apply policy gradient methods to Thompson sampling, allowing us to fine-tune a widely used algorithm with strong theoretical guarantees. Very recently, an independent and contemporaneous pre-print by [37] discovered a similar approach to tuning Thompson sampling.

2. **We provide and analyze multiple gradient estimators for sampling policies.**

   As in the broader application of policy gradient for reinforcement learning, there are multiple possible gradient estimators possible, through different choices of reward metrics and baselines. We derive several novel policy gradient estimators that are specifically tailored to Bayesian bandit problems. We are able to compare their variance theoretically and empirically.

3. **We computationally demonstrate the benefits of our approach.**

   Through simple numerical experiments, we provide a compelling proof of concept. Policy search produces policies that correct for shortcomings of Thompson sampling in short horizon problems or problems with large discrepancies between the variances of arm rewards. Perhaps more surprisingly, policy search offers substantial improvements over Thompson sampling even in a canonical long horizon problem to which it is ideally suited. We also compare against optimistic Gittins indices [38], information directed sampling [39], and Bayesian upper confidence bound algorithms [40], confirming that direct policy search on top of Thompson sampling produces state of the art results for widely studied problem settings. In the future, we hope to extend the numerical experiments beyond problems with independent arms.

3.2 Model

We consider a multi-armed bandit problem in a Bayesian setting.

Rewards. Let $\mathcal{A}$ be the set of arms, possibly infinite, among which the decision maker (DM) can select at each time $t = 1, \ldots, T$. When the DM pulls an arm $a \in \mathcal{A}$ at time t, they earn a random reward $R_{a,t}$ drawn from a distribution $\mathcal{R}$ that is parameterized by *model parameters* $\theta \in \Theta$, i.e.,

$$R_{a,t}|\theta \sim \mathcal{R}(\theta, a). \tag{3.1}$$

We assume that the rewards $(R_{a,t})_{t\in[T]}$ are conditionally independent and identically distributed[1] given θ, for each $a \in \mathcal{A}$. We further define $\mu\colon \Theta \times \mathcal{A} \to \mathbb{R}$ to be the *mean reward function*, i.e.,

$$\mu(\theta, a) \triangleq \mathbb{E}_{R\sim\mathcal{R}(\theta,a)}\left[R\right]. \tag{3.2}$$

As we consider a Bayesian setting, we view θ as a (multivariate) random variable that is sampled from a *prior distribution*, which we denote by $\mathcal{P}(y_0)$, i.e.,

$$\theta \sim \mathcal{P}(y_0), \tag{3.3}$$

where $y_0 \in \mathcal{Y}$ are the sufficient statistics that parameterize the prior distribution.

Information set. The parameters θ are unknown to the DM, but can be inferred by observing the reward realizations sequentially revealed over time. More precisely, let H_t be the history, or information revealed up to time t (inclusive):

$$H_t \triangleq \left(A_s, R_{A_s,s}\right)_{s\in[t]}. \tag{3.4}$$

[1] One important case where the rewards are *not* i.i.d. is the case of contextual bandits. Here, the reward at time t is given by $R_{a,t}|\theta, x_t \sim \mathcal{R}(\theta, a, x_t)$. Here x_t is the context at time t, and this is a stochastic process that evolves independently of the DM's actions. The framework we develop here can be extended to accommodate contextual cases as well in a straightforward fashion, but in the interest of simplifying the presentation we will not be explicit about this.

This includes the actions taken by the DM and the rewards realized through time t. We assume that the DM has knowledge of both the prior distribution $\mathcal{P}(y_0)$ and the functional form of reward distribution $\mathcal{R}$.

As a Bayesian learner, the DM will update their belief according to Bayes' rule whenever observing a new reward realization, and thus maintain a *posterior distribution* for θ at each time. Without loss of generality,[2] we assume that the posterior is represented as

$$\theta | H_t \ \sim \ \mathcal{P}(Y_t). \tag{3.5}$$

Here, $Y_t \in \mathcal{Y}$ is a random variable that denotes the sufficient statistics of the posterior distribution after observing the history H_t (i.e., after observing the t^{th} reward realization). We set $Y_0 = y_0$.

We will describe the randomness of our stochastic model more explicitly as follows. The DM's policy π is described by a sequence of deterministic mappings $(\pi_t)_{t\in[T]}$. Each mapping $\pi_t : \mathcal{H}_{t-1} \times \mathcal{E} \to \mathcal{A}$ specifies the next action A_t as a function of history $H_{t-1} \in \mathcal{H}_{t-1}$ that is revealed immediately prior to time t, and random noise $\epsilon_t \in \mathcal{E}$ that can be utilized for randomization in the choice of action. Similarly, the reward realization is described by a mapping $r : \Theta \times \mathcal{A} \times \Xi \to \mathbb{R}$ that specifies the next reward realization, where any randomness is generated by the noise variable $\xi_t \in \Xi$. That is,

$$A_t = \pi_t(H_{t-1}, \epsilon_t), \quad R_{a,t} = r(\theta, a, \xi_t). \tag{3.6}$$

We assume, without loss of generality, that the noise random variables $(\epsilon_t)_{t\in[T]}$ are independent and identically distributed, as are $(\xi_t)_{t\in[T]}$.

We define an *instance* or *episode*, denoted by I, as a random variable that encodes all uncertainties in the environment, but not the randomness in the DM's decision rule:

$$I \triangleq \left(\theta, (\xi_t)_{t\in[T]}\right). \tag{3.7}$$

[2]When $\mathcal{P}$ is a conjugate prior of $\mathcal{R}$ and belongs to the exponential family, the sufficient statistics Y_t will admit a compact representation. In other cases, Y_t may represent the entire history, i.e., $Y_t = H_t$.

In other words, given an instance I, we exactly know what rewards will be realized for any given action sequence $a_{1:T} \in \mathcal{A}^T$ committed by the DM. The set of all possible instances is denoted by $\mathcal{I}$.

Objective. The DM aims to earn as much reward as possible in expectation. Given the DM's decision rule π, its *expected total reward*, denoted by $\textsc{Reward}[\pi]$, is defined as

$$\textsc{Reward}[\pi] \triangleq \mathbb{E}\left[\sum_{t=1}^{T} R_{A_t,t}\right], \tag{3.8}$$

where the expectation is taken with respect to the randomness of instance (i.e., the randomness of the parameters θ and the reward realizations) and also any randomness of the choice of actions (if π is a randomized policy).

To better illustrate our setup, we provide an example of a canonical multi-armed bandit problem described with the notation introduced above.

Example 3.2.1 (Gaussian MAB)**.** *Consider a finite number of arms $\mathcal{A} \triangleq \{1, \ldots, K\}$. Each arm a yields normally distributed rewards with an unknown mean θ_a and a known variance σ_a^2, where the prior of θ_a is given by a normal distribution $\mathcal{N}(m_{a,0}, v_{a,0}^2)$. The model parameters are given by the vector $\theta \triangleq (\theta_a)_{a\in[K]}$.*

Each instance takes the form $I \triangleq \big(\theta, (\xi_t)_{t\in[T]}\big)$, where $(\xi_t)_{t\in[T]}$ are i.i.d. standard normal random variables that randomize the reward realizations according to

$$\mu(\theta, a) = \theta_a, \quad r(\theta, a, \xi_t) = \theta_a + \sigma_a \xi_t. \tag{3.9}$$

The posterior for a parameter θ_a after time t is given by a normal distribution $\mathcal{N}(m_{a,t}, v_{a,t}^2)$. Here, the sufficient statistics $m_{a,t}$ and $v_{a,t}^2$ can be computed in a closed-form according to

$$m_{a,t} = v_{a,t}^{-2} \times \left(v_{a,0}^{-2} \cdot m_{a,0} + \sigma_a^{-2} \sum_{s=1}^{t} \mathbb{I}_{\{A_s=a\}} R_{A_s,s}\right), \quad v_{a,t}^2 = \left(v_{a,0}^{-2} + \sigma_a^{-2} \sum_{s=1}^{t} \mathbb{I}_{\{A_s=a\}}\right)^{-1}.$$

As a collection of these sufficient statistics across the arms, $Y_t \triangleq (m_{a,t}, v_{a,t}^2)_{a\in[K]} \in \mathbb{R}^{2K}$, determine

the posterior of the parameters θ given the history H_t.

3.3 Parameterized Thompson sampling

Thompson sampling (TS) [41] is a randomized policy that works as follows. At each time $t = 1, \ldots, T$: (i) the parameters $\tilde{\theta}_t$ are sampled from the posterior distribution $\mathcal{P}(Y_{t-1})$ given all information prior to time t; and (ii) an action is chosen to maximize the expected reward given these sampled parameters $\tilde{\theta}_t$. In other words,

$$\tilde{\theta}_t \sim \mathcal{P}(Y_{t-1}), \quad A_t^{\mathrm{TS}} \leftarrow \operatorname*{argmax}_{a \in \mathcal{A}} \mu(\tilde{\theta}_t, a), \tag{3.10}$$

where $Y_{t-1} \in \mathcal{Y}$ are, as introduced in (3.5), the sufficient statistics describing the posterior distribution of true parameters θ given the history H_{t-1}. As time progresses, the above procedure is repeated, while updating the posterior distribution according to Bayes' rule.

An important characteristic of TS is "probability matching". Under TS, the probability that an arm a is selected at time t is equal to the probability that the arm a is indeed the best one that Bayesian inference predicts, i.e.,

$$\mathbb{P}\big(A_t^{\mathrm{TS}} = a \mid H_{t-1}\big) = \mathbb{P}\left(a = \operatorname*{argmax}_{a' \in \mathcal{A}} \mu(\theta, a') \mid H_{t-1}\right), \quad \forall a \in \mathcal{A}. \tag{3.11}$$

However, probability in (3.11) is difficult to evaluate since it does not admit a closed-form expression in most cases and does not admit feasible policy gradient estimators.

We consider a class of variants of TS where the sampling policy in (3.10) is not the posterior distribution, but instead is some other distribution parameterized by *meta-parameters* $\lambda \in \Lambda \subseteq \mathbb{R}^d$. In other words, given λ, the corresponding *sampling policy* $\mathrm{TS}(\lambda)$ repeats the following at each time t:

$$\tilde{\theta}_t \sim \mathcal{P}_\lambda(H_{t-1}), \quad A_t^{\mathrm{TS}(\lambda)} \leftarrow \operatorname*{argmax}_{a \in \mathcal{A}} \mu(\tilde{\theta}_t, a), \tag{3.12}$$

where $\mathcal{P}_\lambda(H_{t-1})$ is a distribution on the parameter space Θ that has arbitrary dependency on the

meta-parameters λ and the history H_{t-1}. The sampling policy $\mathrm{TS}(\lambda)$ is almost identical to the naïve TS except that it samples the parameters from $\mathcal{P}_\lambda(H_{t-1})$ instead of $\mathcal{P}(Y_{t-1})$. In this way, sampling policies can be viewed as a natural generalization of TS, emitting at each time a randomized pseudo-action $\tilde{\theta}_t$ (choice of parameters) from which a base action is determined, rather than directly emitting a base action.

As we would like to employ policy gradient methods to optimize over the meta-parameters λ, we will assume that the probability density (or mass) function of the distribution $\mathcal{P}_\lambda(H_{t-1})$ is differentiable with respect to λ over its domain Λ, for any realization of history H_{t-1} almost surely. Aside from this, the distribution $\mathcal{P}_\lambda(H_{t-1})$ defining the sampling policy is allowed to be essentially arbitrary. However, in order to illustrate our ideas, consider the following example:

Example 3.3.1 (Posterior reshaping.)**.** *We adopt and generalize the idea proposed in [42], and let the algorithm to sample the parameters from the a reshaped posterior distribution,*

$$\mathcal{P}_\lambda(H_{t-1}) = \mathcal{P}\big(\phi_\lambda(Y_{t-1})\big), \tag{3.13}$$

where $\mathcal{P}(\cdot)$ is the posterior distribution defined in (3.5)*, and $\phi_\lambda\colon \mathcal{Y} \to \mathcal{Y}$ is a differentiable mapping, parameterized by λ, that transforms one set of sufficient statistics to another.*

Posterior reshaping is motivated by several arguments. Compared to the general parameterization (3.12), the posterior reshaping does not parameterize "how to learn". Instead, it parameterizes how to utilize the learned results, while maintaining the Bayesian learning logic as it is. This can significantly reduce the effort required for tuning the meta-parameters. Moreover, its implementation requires a minimal effort once one has already implemented the standard TS. Indeed, when $\phi_\lambda(\cdot)$ is the identity map, posterior reshaping reduces to standard TS. Hence, under appropriate technical assumptions, a local policy gradient search starting at the identity map will be guaranteed to do no worse than standard TS.

As discussed in the introduction, the standard TS suffers from the over-exploration, for example when the time horizon is short relative to the number of arms. Posterior reshaping can naturally

address this, by reducing uncertainty in the sampling distribution. As an extreme example, consider a situation where we are given a single time period (i.e., $T = 1$). The optimal policy is myopic — the optimal action is to pick the arm with the largest prior mean — which can be implemented by reshaping the posterior distribution to concentrate on the prior mean. Such posterior concentration also appears in the work of [43]. The IRS.FH policy they suggest is posterior reshaping sampling policy.

Furthermore, it is possible that TS with the correct model parameters (e.g., the specification of the prior distribution or the reward distribution) may not be optimal for the performance of the algorithm. Within the framework of posterior reshaping, we can find the set of model parameters that are empirically tuned for performance so as to outperform to the one with the correct values.[3]

A concrete examples of posterior reshaping in the Gaussian case can be developed as follows:

Example 3.3.2 (Posterior reshaping for Gaussian MAB)**.** *Using the notation of Example 3.2.1, standard* TS *in Gaussian MAB samples the parameters* $\tilde{\theta}_t = (\tilde{\theta}_{a,t})_{a\in\mathcal{A}}$ *according to*

$$\tilde{\theta}_{a,t} \sim \mathcal{N}\left(\frac{m_{a,0} + \frac{v_{a,0}^2}{\sigma_a^2}\cdot S_{a,t-1}}{1 + \frac{v_{a,0}^2}{\sigma_a^2}\cdot N_{a,t-1}},\ \frac{v_{a,0}^2}{1 + \frac{v_{a,0}^2}{\sigma_a^2}\cdot N_{a,t-1}}\right), \tag{3.14}$$

for each $a = 1, \ldots, K$*, where the sufficient statistics are given by* $S_{a,t-1} \triangleq \sum_{s=1}^{t-1} \mathbb{I}_{\{A_s=a\}} R_{A_s,s}$*, and* $N_{a,t-1} \triangleq \sum_{s=1}^{t-1} \mathbb{I}_{\{A_s=a\}}$.

We consider posterior reshaping with meta-parameters $\lambda \triangleq \left(\lambda_a^m, \lambda_a^v, \lambda_a^\sigma, \lambda_a^\gamma\right)_{a\in\mathcal{A}} \in \mathbb{R}^{4K}$ *under which* $\tilde{\theta}_{a,t}$ *is sampled from*

$$\tilde{\theta}_{a,t} \sim \mathcal{N}\left(\frac{\lambda_a^m + \lambda_a^\sigma \cdot S_{a,t-1}}{1 + \lambda_a^\sigma \cdot N_{a,t-1}},\ \frac{\lambda_a^v (1 - t/T)^{\lambda_a^\gamma}}{1 + \lambda_a^\sigma \cdot N_{a,t-1}}\right). \tag{3.15}$$

This policy reduces to standard TS *if we take* $\lambda_a^m = m_{a,0}$ *(prior mean),* $\lambda_a^v = v_{a,0}^2$ *(prior variance),* $\lambda_a^\sigma = v_{a,0}^2/\sigma_a^2$ *(precision ratio), and* $\lambda_a^\gamma = 0$ *(variance decay exponent). The amount of exploration*

[3]Even if the model is mis-specified, the policy gradient method can be applied, but we need to be careful in the choice of gradient estimator in order to avoid a bias in the gradient estimation. See the related discussion in Section 3.4.3.

is controlled by λ_a^v and λ_a^γ. In particular, the term $(1 - t/T)^{\lambda_a^\gamma}$ diminishes exploration near the end of the horizon, where the benefit from exploration is limited.

Note that this parameterization scheme can be represented in the form of (3.13)*, since the sufficient statistics of the realized observations (i.e., $S_{a,t-1}$ and $N_{a,t-1}$) are uniquely determined from those of current posterior distribution (i.e., $m_{a,t-1}$ and $v^2_{a,t-1}$) and prior distribution (i.e., $m_{a,0}$ and $v^2_{a,0}$). Also note that the probability density function is differentiable with respect to λ given that $\lambda_a^v > 0$ and $\lambda_a^\sigma > 0$.*

A more complex example is as follows:

Example 3.3.3 (Deep recurrent neural network parameterization)**.** *One might consider a recurrent neural network (RNN) structure with, at each time t, input $(A_t, R_{A_t,t})$, hidden state $\tilde{Y}_t$ and output being the sampled pseudo-action $\tilde{\theta}_t$. The network would evolve according to*

$$\tilde{Y}_t \leftarrow \phi^Y_{\lambda_Y}(\tilde{Y}_{t-1}, A_t, R_{A_t,t}), \quad \tilde{\theta}_t \sim \mathcal{P}\left(\phi^\theta_{\lambda_\theta}(\tilde{Y}_{t-1})\right).$$

Here, the hidden state $\tilde{Y}_t$ is analogous to a sufficient statistic in that it summarizes the history up to an including time t. Two deep neural networks, $\phi^Y_{\lambda_Y}(\cdot)$ and $\phi^\theta_{\lambda_\theta}(\cdot)$, with weights λ_Y and λ_θ, govern the evolution of the hidden state $\tilde{Y}_t$ and the output $\tilde{\theta}_t$, respectively. The meta-parameters $\lambda \triangleq (\lambda_Y, \lambda_\theta)$ would be optimized with policy gradient methods.

Example 3.3.3 is in the spirit of the approach of [35] and [36], where RNNs were fit with policy search methods, but where the policies output actions. Here, to contrast, the RNN outputs distributional parameters which are then sampled, leveraging on top of the structure of Thompson sampling.

3.4 Policy gradient for Thompson sampling

We aim to search over the meta-parameters $\lambda \in \mathbb{R}^d$ so that the corresponding policy $\mathrm{TS}(\lambda)$ improves over the original TS significantly. For this purpose, we adopt the policy gradient framework,

which applies variants of stochastic gradient ascent to optimize total expected reward. Formally, one can have in mind the iteration,

$$\lambda_{k+1} = \lambda_k + \alpha_k G(\lambda_k, w_k) \tag{3.16}$$

where (α_k) is a step-size sequence, (w_k) is a sequence of i.i.d. random variables, and $G(\lambda_k, w_k)$ is an unbiased gradient, i.e.,

$$\mathbb{E}\left[G(\lambda, w_k)\right] = \nabla_\lambda \textsc{Reward}\big[\mathrm{TS}(\lambda)\big] = \nabla_\lambda \mathbb{E}\left[\sum_{t=1}^{T} R_{A_t^{\mathrm{TS}(\lambda)},t}\right].$$

Typically, the w_k denote the randomness used by a stochastic simulator. For the gradient estimators we use, $w_k \triangleq \left(\epsilon_t^{(k)}, \xi_t^{(k)}\right)_{t\in[T]}$ consists of realizations of the random noise terms that determine the reward realizations and the action selection of a randomized algorithm. In deriving and comparing gradient estimators, we omit the dependence on k. It is worth noting that the iteration (3.16) is meant for illustrative purposes, and other first order stochastic methods, e.g., Adam [44], can also be utilized.

3.4.1 Score function gradient estimation

Most implementations of policy gradient use score function gradient estimation [31]. However, the conventional scheme requires computing $\nabla_\lambda \log \mathbb{P}\big(A_t^{\mathrm{TS}(\lambda)} = a\,\big)$, which is typically intractable since there is no closed-form expression for the distribution of the chosen action.

We circumvent this issue by interpreting the sampled parameters $\tilde{\theta}_t$ as an (pseudo-)action taken by the policy at time t. One can imagine an equivalent bandit environment whose action space is set to the parameter space Θ and the decision maker earns the reward $\tilde{R}_{\tilde{\theta}_t,t} \triangleq R_{A_t,t}$ associated with the arm $A_t \triangleq \operatorname{argmax}_a \mu(\tilde{\theta}_t, a)$ as a result of his decision $\tilde{\theta}_t$. Assume that the sampling distribution $\mathcal{P}_\lambda(H_{t-1})$ under the any H_{t-1} has a probability density function $p_\lambda(\,\cdot\,; H_{t-1})$. Assume as well that

$p_\lambda(\,\cdot\,; H_{t-1})$ is differentiable as a function of λ. This leads to the gradient estimator

$$G \triangleq \sum_{t=1}^{T} S_t \sum_{s=t}^{T} R_{A_s^{\mathrm{TS}(\lambda)},s} \tag{3.17}$$

where

$$S_t \triangleq \nabla_\lambda \log p_\lambda\left(\tilde{\theta}_t\,; H_{t-1}\right) \tag{3.18}$$

denotes the score functions. This form of score function gradient estimator is well known to be unbiased, i.e., $\mathbb{E}[G] = \nabla_\lambda \textsc{Reward}\big[\mathrm{TS}(\lambda)\big]$, which is referred to as the policy gradient theorem in the reinforcement learning literature. Formally, unbiasedness requires technical conditions that allow for the interchange of integrals and derivatives. We refer to [45] for appropriate conditions.

3.4.2 Admissible gradient estimators

The standard gradient estimator (3.17) can be very noisy due to the high variability of reward realizations and random action selections. In this section, we propose a broader, more general class of gradient estimators, and demonstrate that they remain unbiased as long as they satisfy a certain admissibility requirement. Later, we will suggest a specific list of estimators for bandit problems, and provide a theoretical comparison among them in terms of variance reduction.

General representation. With processes $M \triangleq (M_t)_{t\in[T]}$, which we call a *reward metric*, and $B \triangleq (B_t)_{t\in[T]}$, which we call a *baseline*, we define the gradient estimator $G^{M,B}$ as follows:

$$G^{M,B} \triangleq \sum_{t=1}^{T} S_t \cdot (M_t - B_t)\,. \tag{3.19}$$

The reward metric M_t is a random variable that accounts for the sum of rewards that the policy earns on the remaining horizon $t, t+1, \ldots, T$, and B_t is another random variable that represents some benchmark for the rewards over the same period. Note that the estimator (3.17) can be obtained by taking $M_t = \sum_{s=t}^{T} R_{A_s,s}$ and $B_t = 0$.

Admissibility. We state a condition on the reward metric M and the baseline B under which they

induce an unbiased gradient estimator $G^{M,B}$.

Definition 3.4.1 (Admissible reward metric and baseline)**.** *A reward metric M is admissible if for all $t \in [T]$ it is integrable and*

$$\mathbb{E}\left[M_t \middle| H_{t-1}, \tilde{\theta}_t\right] = \mathbb{E}\left[\sum_{s=t}^{T} R_{A_s,s} \middle| H_{t-1}, \tilde{\theta}_t\right]. \tag{3.20}$$

A baseline B is admissible if B_t is integrable and B_t and $\tilde{\theta}_t$ are conditionally independent given H_{t-1} for all $t \in [T]$, i.e.,

$$B_t \perp\!\!\!\perp \tilde{\theta}_t \,\middle|\, H_{t-1}. \tag{3.21}$$

The first condition (3.20) ensures that a risk-neutral decision maker would not differentiate between M_t and the sum of future rewards when deciding the next action. These two measures have the same expectation given any history (which corresponds to the state in dynamic programming terms) and any pseudo-action $\tilde{\theta}_t$. The second condition (3.21) ensures that the decision-maker does not need to take baseline into consideration when making a decision, since the baseline is independent of the next pseudo-action $\tilde{\theta}_t$.

The above interpretation implies that the substitution of reward metric (from $\sum_{s=t}^{T} R_{A_s,s}$ to M_t) and the presence of baseline B_t do not affect the DM's decision at each time t, as long as they satisfy the admissibility conditions (3.20)–(3.21). Therefore, we can infer that the generalized gradient estimator (3.19) is equal in expectation to the standard one (3.17), which is proved formally in the following theorem.

Theorem 3.4.1 (Unbiasedness of gradient estimator)**.** *If the reward metric M and the baseline B are admissible, then $\mathbb{E}[G^{M,B}] = \mathbb{E}[G]$.*

Proof. Note that S_t is measurable with respect to $\sigma\left(H_{t-1}, \tilde{\theta}_t\right)$ and

$$\mathbb{E}\left[S_t \middle| H_{t-1}\right] = 0, \tag{3.22}$$

due to the property of the score function. By the condition (3.20), we obtain

$$\begin{aligned}
\mathbb{E}\left[S_t M_t\right] &= \mathbb{E}\left[\mathbb{E}\left(S_t M_t \,\middle|H_{t-1}, \tilde{\theta}_t\right)\right] \\
&= \mathbb{E}\left[S_t \times \mathbb{E}\left(M_t \,\middle|H_{t-1}, \tilde{\theta}_t\right)\right] \\
&= \mathbb{E}\left[S_t \times \mathbb{E}\left(\sum_{s=t}^{T} R_{A_s,s} \,\middle|\, H_{t-1}, \tilde{\theta}_t\right)\right] \\
&= \mathbb{E}\left[\mathbb{E}\left(S_t \times \sum_{s=t}^{T} R_{A_s,s} \,\middle|\, H_{t-1}, \tilde{\theta}_t\right)\right] \\
&= \mathbb{E}\left[S_t \times \sum_{s=t}^{T} R_{A_s,s}\right].
\end{aligned}$$

By the condition (3.21), we further obtain

$$\mathbb{E}\left[S_t B_t\right] = \mathbb{E}\left[\mathbb{E}\left(S_t B_t | H_{t-1}\right)\right] = \mathbb{E}\left[\mathbb{E}\left(S_t | H_{t-1}\right) \times \mathbb{E}\left(B_t | H_{t-1}\right)\right] = 0.$$

Combining these results, we deduce that

$$\mathbb{E}\left[G^{M,B}\right] = \mathbb{E}\left[\sum_{t=1}^{T} S_t \times (M_t - B_t)\right] = \mathbb{E}\left[\sum_{t=1}^{T} S_t \sum_{s=t}^{T} R_{A_s,s}\right] = \mathbb{E}[G],$$

which concludes the proof. □

One particularly interesting class of reward metrics and baselines are those which are time separable:

Example 3.4.1 (Time separable reward metric and baseline)**.** *Consider*

$$M_t \triangleq \sum_{s=t}^{T} \hat{r}_s(A_{1:s}, I), \quad B_t \triangleq b_t(A_{1:t-1}, I),$$

where $\hat{r}_t \colon \mathcal{A}^t \times \mathcal{I} \to \mathbb{R}$ *and* $b_t \colon \mathcal{A}^{t-1} \times \mathcal{I} \to \mathbb{R}$ *are the deterministic functions that satisfy*

$$\mathbb{E}\left[\hat{r}_t(a_{1:t}, I) \mid H_{t-1}, A_{1:t-1} = a_{1:t-1}\right] = \mathbb{E}\left[r(\theta, a_t, \xi_t) \mid H_{t-1}, A_{1:t-1} = a_{1:t-1}\right], \quad \forall\, a_{1:t} \in \mathcal{A}^t,\ t \in [T].$$

Then, the reward metric $M \triangleq (M_t)_{t\in[T]}$ *and the baseline* $B \triangleq (B_t)_{t\in[T]}$ *are admissible.*

We remark that the baseline is allowed to be *instance-dependent*, meaning that it can depend on instance I defined in (3.7) that determines the realizations of rewards and the true parameter θ. This is a considerable generalization of the literature in which baselines are typically chosen as a deterministic function of state [46]. The use of common randomness [47] in the the baseline and the reward metric can reduce variance, especially when most variation in observed algorithm performance is driven by different realizations of the problem instance rather than differences in the choice of meta parameter.

3.4.3 Reward metrics and baselines

We suggest a specific series of reward metrics and baselines that are admissible for Bayesian bandit problems. A number of these take the time separable form of Example 3.4.1.

Reward metrics. The followings are possible choices for reward metric:

1. The observed reward $M_t^{\text{obs}} \triangleq \sum_{s=t}^{T} R_{A_s,s}$.

2. The mean reward $M_t^{\text{mean}} \triangleq \mathbb{E}\left[\sum_{s=t}^{T} R_{A_s,s} \middle| \theta, A_{t:T}\right] = \sum_{s=t}^{T} \mu(\theta, A_s)$.

3. (MAB with independent arms only) The finite-sample mean-reward estimate

$$M_t^{\text{fin}} \triangleq \sum_{s=t}^{T} \hat{\mu}_{I,t}(A_s),$$

   where $\hat{\mu}_{I,t}(a) \triangleq \mathbb{E}\left[\mu(\theta, a) \,|H_T, A_s = a, \forall s = t, \ldots, T\right]$ that indicates the best estimate for the mean reward of an arm a that the DM can infer through a finite number of observations.[4]

4. The posterior mean $M_t^{\text{Bayes}} \triangleq \sum_{s=t}^{T} \mathbb{E}\left[\mu(\theta, A_s) \,|H_{s-1}, A_s\right]$.

[4]This metric is valid only when the arms and their associated priors are independent. Using the notation of (3.6), we further need to assume that the noise variable takes the form $\xi_t \triangleq (\xi_{a,t})_{a\in\mathcal{A}}$ where $\xi_{a,t}$ independent across a. In order for the DM to retrieve maximal information about a particular arm a, it is required to pull the arm a throughout the entire rest of the horizon (i.e., $A_s = a, \forall s = t, \ldots, T$). The metric $\hat{\mu}_{I,t}(a)$ represents the mean reward estimate that the DM will have in this scenario, which also has a dependency on the instance I.

5. The state-action Q-function

$$M_t^Q \triangleq \mathbb{E}\left[\sum_{s=t}^{T} R_{A_s,s} \middle| H_{t-1}, A_t\right] = \mathbb{E}\left[\sum_{s=t}^{T} \mu(\theta, A_s) \middle| H_{t-1}, A_t\right].$$

Recall that $\mu(\theta, a)$ is the mean reward function, defined in (3.2), that is a deterministic function representing the expected reward of an arm a given the parameters θ.

These metrics differ in the information set on which the conditional expectation of the sum of future rewards, $\sum_{s=t}^{T} R_{A_s,s}$, is taken. The main motivation for deriving this series of metrics is "Rao–Blackwellization," i.e., integrating out some of the randomness in the future reward realizations and the future action selections. More specifically, the metric M_t^{mean} is obtained from M_t^{obs} by integrating out the randomness of immediate reward realization while maintaining the dependency on the (random) parameters θ and the (random) action sequence $A_{t:T}$. The metric M_t^{fin} is motivated from the fact that knowing the true parameters θ is as informative as having an infinite number of observations for each arm, and improves over M_t^{mean} by taking into account how much the DM can learn about θ with a finite number observations (i.e., by integrating out the uncertainties in θ that cannot be identified). Next, under the metric M_t^{Bayes}, the DM earns the expected reward given the posterior distribution at each time, which averages out the uncertainties in θ at each time step. Finally, the metric M_t^Q represents the Q-value of the given policy, i.e., the expected future reward of the policy at a given state (history) and an action (arm), which averages out the all uncertainties that arise after taking the action A_t.

We remark that these reward metrics are mostly taken from [43]. While the reward metrics M_t^{obs} and M_t^Q are applicable for the general Markov decision processes, the other three metrics M_t^{mean}, M_t^{fin} and M_t^{Bayes} are valid only for bandit problems, and in particular, M_t^{fin} and M_t^{Bayes} are valid only in a Bayesian setting. In addition, accurate computation of M_t^Q typically requires averaging over many Monte Carlo simulations (e.g., roll-outs) which may be computationally expensive.

Baselines. We further provide a list of baselines as follows:

1. The null baseline $B_t^{\text{null}} \triangleq 0$.

2. The oracle performance $B_t^{\text{oracle}} \triangleq M_t^\star$ where $M_t^\star$ is the reward (measured with the corresponding reward metric) that the action sequence $A_t^\star = \text{argmax}_{a\in\mathcal{A}}\, \mu(\theta, a)$ achieves in the *same instance*. For example, in a combination with M_t^{mean}, we obtain $B_t^{\text{oracle}} = \sum_{s=t}^{T} \max_{a\in\mathcal{A}} \mu(\theta, a)$.

3. The self-play baseline $B_t^{\text{self}} \triangleq \tilde{M}_t$ where $\tilde{M}_t$ is the reward (measured with the corresponding reward metric) that an independent run of the same algorithm achieves in the same instance. For example, in a combination with M_t^{mean}, we obtain $B_t^{\text{self}} = \sum_{s=t}^{T} \mu(\theta, \tilde{A}_s)$ where $\tilde{A}_{1:T}$ is the action sequence taken in the independent run.

4. The value function $B_t^V \triangleq \mathbb{E}\left[\sum_{s=t}^{T} R_{A_s,s} \middle| H_{t-1}\right]$.

As proven in Theorem 3.4.1, each of these baselines can be used in a combination with any of the reward metrics listed above. The baseline B_t^{oracle} is an instance-dependent measure that represents the performance of the omniscient policy that knows the values of true parameters θ. Given that M_t^{mean} is chosen as a coupled reward metric, the gap $B_t^{\text{oracle}} - M_t^{\text{mean}}$ reduces to the "regret" which is a measure of suboptimality that has been widely used in bandit studies. This choice of baseline is natural when we expect adaptive algorithm to have small average regret. It can be less effective in problems with a short time horizon, where the reward earned by an oracle is not an attainable baseline.

The baseline B_t^{self} utilizes an independent run of the same randomized policy under the same instance. The idea of self-play was adopted from [36] while we make a generalization regarding the choice of reward metric and provide a formal proof of its validity. It effectively centers the reward metric, i.e., $\mathbb{E}[M_t - B_t^{\text{self}}] = 0$, which helps stabilize gradient estimates. In our numerical experiments, B_t^{self} shows an impressive performance across the different settings, though it effectively requires the computational effort of running twice as many simulations.

Finally, the baseline B_t^V is constructed analogously to the reward metric M_t^Q, and it represents the average performance of the given policy at the given state. In a combination with M_t^Q, the gap $M_t^Q - B_t^V$ measures the relative benefit of the chosen action compared to the average, which is also known as the advantage function. Like M_t^Q, however, this baseline does not have a closed-form

expression. The baseline B_t^V can be understood as averaging the result of B_t^{self} (applied with the posterior mean reward metric) across many independent runs of the algorithm. The randomized baseline B_t^{self} has higher variance, but can be calculated at much lower computational cost.

Implementation issues. If we are equipped with a simulator that can generate instances with full information, it is straightforward to compute the reward metrics and the baselines listed above (apart from the computational efficiency). If we are running the algorithm in the real world situation, however, we may not be able to identify their values as we do not have an access to unrevealed information such as true model parameters θ. Nevertheless, in the Bayesian setting, we can overcome this issue by sampling the unobserved variables at the end of an episode: For example, after completing an episode, we can sample $\tilde{\theta} \sim \mathcal{P}(Y_T)$ as if we perform one more step of TS, and plug them into the formulas for reward metric or baseline. This is valid since $\tilde{\theta}$ is identically distributed with the true parameters θ given the observations revealed in that episode, by the virtue of posterior distribution, and therefore the resulting gradient estimates $\tilde{G}$ will also be identically distributed with the true one G.

Note that any model mis-specification can lead to a bias of the gradient estimator. More specifically, if the prior distribution or the reward distribution is mis-specified (e.g., the value of noise variance σ_a^2 is incorrect in Example 3.2.1), the reward metrics M_t^{fin} and M_t^{Bayes} will result in biased estimates. If the mean reward function $\mu(\cdot, \cdot)$ is incorrect, furthermore, all the reward metrics other than M_t^{obs} will suffer from the bias. We expect that the users can determine whether there is an bias during the training process and adopt a more robust metric if needed.

3.4.4 Variance comparison

The variance of a gradient estimator is a crucial factor for the performance of policy gradient. In this section, we provide an analysis than can provide theoretical comparisons between estimators of the form (3.19), including many of the estimators in Section 3.4.3.

To begin, note that for an admissible estimator of the form (3.19), we have

$$\mathbb{E}\left[G\right] = \mathbb{E}\left[G^{M,B}\right] = \mathbb{E}\left[\sum_{t=1}^{T} S_t \cdot (M_t - B_t)\right] = T \times \mathbb{E}\left[\frac{1}{T}\sum_{t=1}^{T} G_t^{M,B}\right] = \mathbb{E}\left[T \times G_\tau^{M,B}\right],$$

where $G_t^{M,B} \triangleq S_t \cdot (M_t - B_t)$, and $\tau \in [T]$ is a random time index that is independently and uniformly distributed. Thus, given any admissible estimator $G^{M,B}$, we can construct a related *single time* estimator $T \times G_\tau^{M,B}$ that is also unbiased. Loosely speaking, this estimator estimates the gradient based on a the impact of an action taken at a single, randomly chosen decision epoch τ, rather than considering all decision epochs. Moreover, the simpler, single time estimator is more amenable to analysis.

In the next theorem, we further provide a comparison between two single time gradient estimators in terms of the variance they induce. For two square symmetric matrices A and B, we say $A \preceq B$ if and only if $B - A$ is a positive semi-definite matrix. This gives a partial ordering of symmetric matrices.

Theorem 3.4.2 (Variance reduction)**.** *Consider two reward metric and baseline pairs* $(\underline{M}_t, \underline{B}_t)$ *and* $(\overline{M}_t, \overline{B}_t)$ *that satisfy*

$$\underline{M}_t - \underline{B}_t = \mathbb{E}\left[\overline{M}_t - \overline{B}_t \middle| \mathcal{G}_t\right], \quad \forall t \in [T], \tag{3.23}$$

for some $\mathcal{G}_t \supseteq \sigma(H_{t-1}, \tilde{\theta}_t)$*. Let* $\underline{G}_\tau$ *and* $\overline{G}_\tau$ *be corresponding single time gradient estimators, respectively, i.e.,*

$$\underline{G}_\tau \triangleq S_\tau \cdot \left(\underline{M}_\tau - \underline{B}_\tau\right), \quad \overline{G}_\tau \triangleq S_\tau \cdot \left(\overline{M}_\tau - \overline{B}_\tau\right).$$

Then, $\overline{G}_\tau$ *exhibits a smaller variance than* $\underline{G}_\tau$*, in the sense that*

$$\mathrm{Cov}[\underline{G}_\tau] \preceq \mathrm{Cov}[\overline{G}_\tau]. \tag{3.24}$$

Proof. Fix $t \in [T]$. By the law of total covariance and conditioning on $\mathcal{G}_t$,

$$\begin{aligned}
\mathrm{Cov}[\overline{G}_t] &= \mathrm{Cov}\left[\mathbb{E}(\overline{G}_t|\mathcal{G}_t)\right] + \mathbb{E}\left[\mathrm{Cov}\left(\overline{G}_t|\mathcal{G}_t\right)\right] \\
&\geq \mathrm{Cov}\left[\mathbb{E}(\overline{G}_t|\mathcal{G}_t)\right] \\
&= \mathrm{Cov}\left[\mathbb{E}\left(S_t \cdot (\overline{M}_t - \overline{B}_t)\middle|\mathcal{G}_t\right)\right] \\
&= \mathrm{Cov}\left[S_t \cdot \mathbb{E}\left(\overline{M}_t - \overline{B}_t\middle|\mathcal{G}_t\right)\right] \\
&= \mathrm{Cov}\left[S_t \cdot (\underline{M}_t - \underline{B}_t)\right] \\
&= \mathrm{Cov}[\underline{G}_t],
\end{aligned} \tag{3.25}$$

where the inequality in the second step follows from the fact that every covariance matrix is positive semi-definite.

Now, note that

$$\mathbb{E}\left[\overline{G}_\tau\middle|\tau\right] = \mathbb{E}\left[S_\tau \cdot \mathbb{E}\left[\overline{M}_\tau - \overline{B}_\tau\middle|\mathcal{G}_\tau\right]\middle|\tau\right] = \mathbb{E}\left[S_\tau \cdot (\underline{M}_\tau - \underline{B}_\tau)\middle|\tau\right] = \mathbb{E}\left[\underline{G}_\tau\middle|\tau\right].$$

Then, applying the law of total covariance again, this time conditioning on τ,

$$\begin{aligned}
\mathrm{Cov}[\overline{G}_\tau] &= \mathrm{Cov}\left[\mathbb{E}(\overline{G}_\tau|\tau)\right] + \mathbb{E}\left[\mathrm{Cov}\left(\overline{G}_\tau|\tau\right)\right] \\
&= \mathrm{Cov}\left[\mathbb{E}(\underline{G}_\tau|\tau)\right] + \mathbb{E}\left[\mathrm{Cov}\left(\overline{G}_\tau|\tau\right)\right] \\
&\geq \mathrm{Cov}\left[\mathbb{E}(\underline{G}_\tau|\tau)\right] + \mathbb{E}\left[\mathrm{Cov}\left(\underline{G}_\tau|\tau\right)\right] \\
&= \mathrm{Cov}[\underline{G}_\tau].
\end{aligned}$$

Here, the inequality follows from (3.25). This concludes the proof. □

Theorem 3.4.2 provides a pairwise comparison of two single time gradient estimators (i.e., $\overline{G}_\tau$ and $\underline{G}_\tau$), when their reward metrics and baselines are related by (3.23). Ideally we would like a comparison between the variance of the original gradient estimators (i.e., $\mathrm{Cov}\left[\sum_{t=1}^T \underline{G}_t\right]$ and $\mathrm{Cov}\left[\sum_{t=1}^T \overline{G}_t\right]$). However, this is challenging due to the interdependence across time between the

score functions and the reward metrics. Nevertheless, we believe that Theorem 3.4.2 is informative, and the ordering it implies is consistent with the numerical performance results we will see in Section 3.5.

Theorem 3.4.2 implies that the reward metric based on the smaller information set (i.e., through more averaging) produces a more precise gradient estimator than one based on the larger information set (i.e., with less averaging). This is the same insight that drives the Rao-Blackwell theorem.

In the development of the reward metrics in Section 3.4.3, we have argued that some reward metrics are motivated from the others via Rao-Blackwellization. In fact, the relationship (3.23) holds among the reward metrics M_t^{obs}, M_t^{mean}, M_t^{fin}, and M_t^{Q} (not including M_t^{Bayes}). Indeed, an application of Theorem 3.4.2 immediately yields the following ordering among the gradient estimators:

Corollary 3.4.1.

$$\text{Cov}[G_\tau^{M^{\text{obs}},B}] \geq \text{Cov}[G_\tau^{M^{\text{mean}},B}] \geq \text{Cov}[G_\tau^{M^{\text{fin}},B}] \geq \text{Cov}[G_\tau^{M^{Q},B}],$$

for any choice of baseline B from B^{null}, B^{oracle}, B^{self}, and B^{V}. Here, note that the baselines B_t^{oracle} and B_t^{self} require a coupled reward metric. We assume they are coupled to the corresponding reward metric in use in each estimator.

3.5 Numerical experiments

In this section, we report the simulation results of the policy gradient optimization of Thompson sampling. We aim to illustrate the flexibility of our proposed framework as a meta-learning platform for bandit tasks, compare the gradient estimators with different choices of reward metric and baseline, and highlight the performance of optimized sampling policies in a comparison with the other state-of-the-art algorithms.

Setup. We consider Gaussian multi-armed bandit (MAB) problems, introduced in Example 3.2.1, for which we implement TS with parameterized posterior reshaping, described in Example 3.3.2.

To highlight the improvement over the naïve TS, our experiments include the following configurations:

1. Gaussian MAB with 10 arms ($K = 10$) and 500 time periods ($T = 500$), where all arms have the same prior distribution and the same noise variance. This is a typical setting that has been studied in many prior works.

2. Gaussian MAB with heteroscedastic reward distributions, where we are given 5 arms ($K = 5$) with very different noise variances and 50 time periods ($T = 50$). Since each arm requires a different amount of effort to learn its unknown mean, it is important to incorporate information about the noise variances into the decision making, which standard TS does not do.

3. Gaussian MAB with an excessive number of arms, where we are given 20 arms ($K = 20$) with identical priors and 20 time periods ($T = 20$). In this setup, there is no hope of discovering the true optimal arm. Nevertheless, standard TS continues to select arms that have never been tried throughout the entire time horizon, which is very wasteful exploration.

Note that in all of these settings we have adopted the same parameterization of TS. This is to verify that our proposed framework achieves the goal of meta-learning: The policy gradient procedure finds the choice of meta-parameters $\lambda \in \mathbb{R}^{4K}$ from Example 3.3.2 that is optimized for each of the bandit settings, resulting in the algorithm that is trained to exploit the structure in each setting and performs no worse than the standard version of TS. We highlighted that the optimized behavior differs substantially across settings and at times differs substantially from TS.

Training. We implement the policy gradient algorithm based on the gradient estimator (3.19) with different combinations of reward metric M and baseline B. In each policy gradient iteration, we compute the batch gradient, i.e., the average gradient measured across a set of independently generated bandit instances, where the batch size (the number of instances) ranges from 1,000 to 5,000 across the settings. The Adam optimizer [44] is used to perform the gradient ascent steps.

The random generation of Gaussian MAB instances is done according to the model described in Example 3.2.1. To facilitate an accurate comparison between the gradient estimators, the estimators share all the randomness in the instance generation and the random action selection. That is, in the notation of (3.6), the same realizations of noise variables (ϵ_t) and (ξ_t) are used for the simulation of different policy gradient estimators.

Evaluation. As a suboptimality measure of a bandit algorithm, we utilize the (Bayesian) regret defined as follows:

$$\textsc{Regret}(\pi) \triangleq \mathbb{E}\left[\sum_{t=1}^{T} \max_{a\in\mathcal{A}}\{\mu(\theta, a)\} - \mu(\theta, A_t)\right], \tag{3.26}$$

which is measured via sample average approximation in our simulation. When computing the gradient estimator during the training process, we obtain as a side product the regret that the algorithm incurs in each training batch, and we report this trajectory of regret as a learning curve of the policy gradient optimization. We naturally expect that the regret decreases as training proceeds. Finally, we measure the regret of the trained policies (and the other bandit algorithms listed below) on the evaluation batch, which is a set of instances generated independently of the training batches. As done in training, the same set of instances are used for evaluating all the policies so as to facilitate accurate comparisons among them.

Competing bandit algorithms. We consider the state-of-the-art bandit algorithms that are suitable for a Bayesian setting: the Bayesian upper confidence bound [40] (BAYES-UCB, with a quantile of $1 - \frac{1}{t}$), information-directed sampling [39] (IDS), and the optimistic Gittins index[5] [38] (OGI). We compare the performance of the trained TS policies with these algorithms.

Implementation. All the code is written in Python, and the training module is implemented using Tensorflow to utilize the automatic gradient calculation and the Adam optimizer. We use 64-bits floating numbers for computation of gradient estimator.

[5]There are two free parameters in OGI. We use a one-step look-ahead and a discount factor of $\gamma_t = 1 - \frac{1}{t}$, which was the primary focus of [38].

3.5.1 Gaussian MAB in a standard setting ($K = 10, T = 500$)

We first report the result for Gaussian MAB with 10 arms and 500 time periods. More specifically, we are given ten independent arms with identical prior distributions: For each arm $a = 1, \ldots, K$ and time $t = 1, \ldots, T$, we assume that

$$\theta_a \sim \mathcal{N}(0, 1^2), \quad R_{a,t}|\theta_a \sim \mathcal{N}(\theta_a, 1^2). \tag{3.27}$$

This setup has been also examined in the prior literature [38, 39].

For policy gradient optimization of TS with parameterized posterior reshaping, we adopt the various combinations of reward metric M and baseline B for the gradient estimator $G^{M,B}$. The initial values for the meta-parameters λ are chosen in the way that the corresponding policy is identical to the standard TS. The training batch size is set to 5,000 and the learning rate for Adam optimizer is set to 0.01.

Figure 3.1 shows the learning curves obtained in our simulated training, and Table 3.1 reports the performance of the trained TS policies as well as the other algorithms being compared. In every combination of reward metric and baseline, we observe a steady improvement in performance over the course of the training process (starting from the standard TS). The training performance largely depends on the choice of baseline: with baseline B^{oracle} or B^{self} the algorithm shows an impressive progress, catching the state-of-the-art algorithms within 300 policy gradient iterations and ending up with policies that improve over the standard TS by 23% in terms of regret.

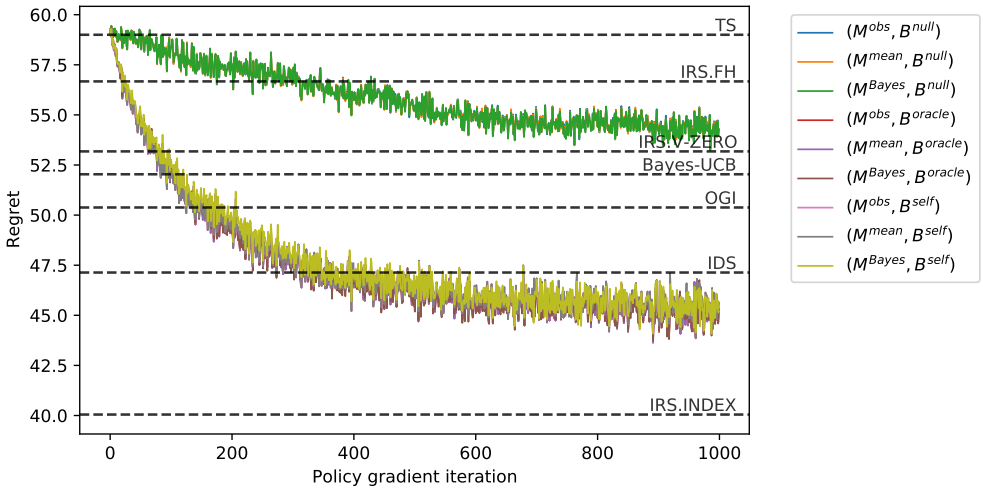

Figure 3.1: Learning curves of parameterized Thompson sampling trained for Gaussian MAB with 10 arms and 500 time periods. A curve shows the progress of policy gradient optimization based on a particular choice of reward metric M and baseline B for the gradient estimator $G^{M,B}$, defined in (3.19). The k^{th} data point in each curve reports the average regret on the k^{th} training batch, which contains 5,000 independent instances. The dashed horizontal lines represent the performance of the other algorithms measured with the evaluation batch containing 20,000 instances (also reported in Table 3.1).

Algorithm	Reward metric	Baseline	$\mathrm{Tr}(\mathrm{Cov}[G^{M,B}])$ ($\times 10^6$)	Regret (s.e.)
Trained TS	M^{obs}	B^{null}	3417.37	54.416 (0.208)
	M^{mean}	B^{null}	3410.17	54.251 (0.197)
	M^{Bayes}	B^{null}	3402.82	54.454 (0.207)
	M^{obs}	B^{oracle}	7.84	45.699 (0.306)
	M^{mean}	B^{oracle}	7.84	**45.360** (0.299)
	M^{Bayes}	B^{oracle}	7.97	45.335 (0.306)
	M^{obs}	B^{self}	4.50	45.913 (0.306)
	M^{mean}	B^{self}	4.50	45.409 (0.295)
	M^{Bayes}	B^{self}	6.51	46.331 (0.318)
Naïve TS	–	–	–	58.999 (0.191)
BAYES-UCB	–	–	–	52.038 (0.186)
OGI	–	–	–	50.381 (0.348)
IDS	–	–	–	47.135 (0.335)
IRS.FH	–	–	–	56.672 (0.180)
IRS.V-ZERO	–	–	–	53.179 (0.187)
IRS.INDEX	–	–	–	**40.048** (0.251)

Table 3.1: Performance of the algorithms for Gaussian MAB with 10 arms and 500 time periods. Each trained TS uses the meta-parameters that are obtained from training procedure, i.e., the ones found at the end of 1,000 iterations of batched policy gradient ascent (Figure 3.1). The performance is measured in regret, defined in (3.26), and computed via sample average approximation over 20,000 independent instances, and reported with the standard error. The best results are emphasized with bold letters.

3.5.2 Gaussian MAB with heteroscedastic arms ($K = 5, T = 50$)

We now explore a different configuration of the Gaussian MAB under which the naïve TS performs particularly poorly. We consider five arms that have significantly different noise variances:

For each arm $a = 1, \ldots, 5$ and time $t = 1, \ldots, 50$, we assume that

$$\theta_a \sim \mathcal{N}(0, 1^2), \quad R_{a,t}|\theta_a \sim \mathcal{N}(\theta_a, \sigma_a^2), \quad \text{where} \quad \sigma_{1:5}^2 \triangleq (0.1,\ 0.4,\ 1,\ 4,\ 10). \tag{3.28}$$

Note that it is crucial for the algorithms to consider the heterogeneity in the reward variances since the variance σ_a^2 determines how much the decision maker can learn about the unknown mean reward θ_a within a finite number of observations: in order for the posterior distribution to concentrate so as to have the standard deviation of 0.1, for example, a single observations is enough for arm 1 whereas 100 and 10,000 observations are required for arm 3 and arm 5, respectively. This is especially important when the time horizon is short, as in this case.

We use the training batches of size 1,000 for gradient estimation, and the Adam optimizer with learning rate of 0.05 for policy gradient, and the evaluation batch of size 10,000 for evaluation. While every combination of reward metric and baseline shows a very stable progress throughout the policy gradient procedure, as shown in Figure 3.2, we observe that the baseline B^{self} works slightly better than the baseline B^{oracle}, and so does B^{oracle} than B^{null}.

The evaluation results are shown in Table 3.2. We immediately observe that naïve TS and BAYES-UCB particularly perform poorly as they make decisions based only on the posterior at each moment without incorporating the noise variances into consideration. As the results show, by optimizing posterior reshaping parameters, we can make TS to trade off exploitation and exploration much more precisely, so that we can achieve a surprising improvement over TS by 35%–40%.

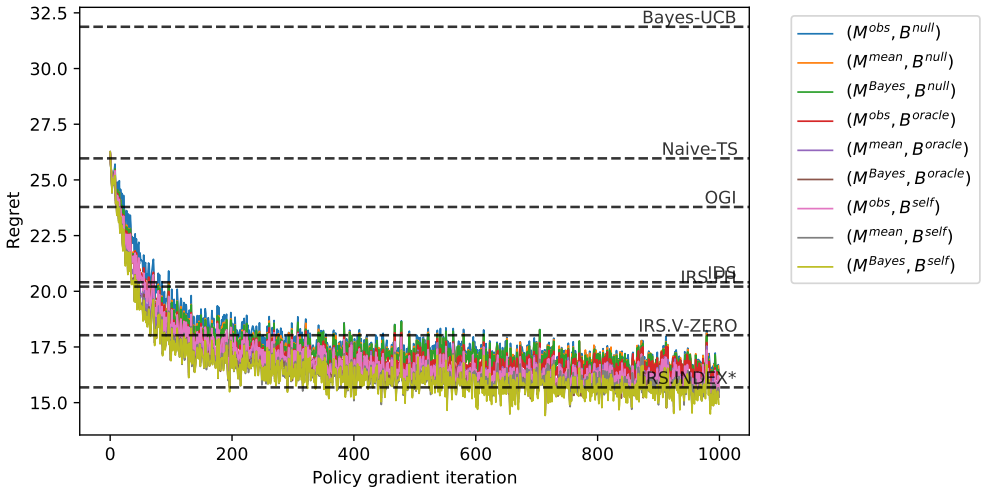

Figure 3.2: Learning curves of parameterized Thompson sampling trained for Gaussian MAB with heteroscedastic arms ($K = 5, T = 50, \sigma^2_{1:5} = (0.1, 0.4, 1, 4, 10)$). A curve shows the progress of policy gradient optimization based on a particular choice of reward metric M and baseline B for the gradient estimator $G^{M,B}$, defined in (3.19). The k^{th} data point in each curve reports the average regret on the k^{th} training batch, which contains 1,000 independent instances. The dashed horizontal lines represent the performance of the other algorithms measured with the evaluation batch containing 10,000 instances (also reported in Table 3.2).

Algorithm	Reward metric	Baseline	$\mathrm{Cov}[G^{M,B}]$ ($\times 10^4$)	Regret (s.e.)
Trained TS	M^{obs}	B^{null}	133.85	16.654 (0.200)
	M^{mean}	B^{null}	75.33	16.723 (0.201)
	M^{Bayes}	B^{null}	66.20	16.546 (0.200)
	M^{obs}	B^{oracle}	78.76	16.315 (0.201)
	M^{mean}	B^{oracle}	22.73	15.864 (0.201)
	M^{Bayes}	B^{oracle}	9.75	15.693 (0.194)
	M^{obs}	B^{self}	59.97	15.885 (0.199)
	M^{mean}	B^{self}	49.15	15.417 (0.194)
	M^{Bayes}	B^{self}	38.99	**15.313** (0.194)
NAIVE-TS	–	–	–	25.967 (0.158)
BAYES-UCB	–	–	–	31.875 (0.256)
OGI	–	–	–	23.785 (0.227)
IDS	–	–	–	20.405 (0.205)
IRS.FH	–	–	–	20.209 (0.169)
IRS.V-ZERO	–	–	–	18.027 (0.175)
IRS.INDEX*	–	–	–	**15.685** (0.200)

Table 3.2: Performance of the algorithms for Gaussian MAB with heteroscedastic arms. Each trained TS uses the meta-parameters found at the end of 1,000 iterations of batched policy gradient ascent (Figure 3.2). The performance is measured in regret, defined in (3.26), and computed via sample average approximation over 10,000 independent instances, and reported with the standard error. The best results are emphasized with bold letters.

3.5.3 Gaussian MAB with an excessive number of arms ($K = 20, T = 20$)

We finally investigate Gaussian MAB with an excessive number of arms, i.e., too many arms compared to the length of time horizon. More specifically, we consider 20 arms and 20 time

periods: For each arm $a = 1, \ldots, 20$ and time $t = 1, \ldots, 20$, we assume that

$$\theta_a \sim \mathcal{N}(0, 1^2), \quad R_{a,t}|\theta_a \sim \mathcal{N}(\theta_a, 1^2). \tag{3.29}$$

This setup is motivated from [48] in which the authors posit an extreme example where TS faces an infinite number of arms with identical priors. In such an example, TS keeps pulling a new arm throughout the entire process, since with zero probability the same arm gets the largest sampled mean $\tilde{\theta}_{a,t}$ more than once: As a result, TS does not utilize any information obtained from the past pulls, and always earns the prior mean $\mathbb{E}[\theta_a]$ in expectation at each time. We aim to see whether TS can resolve this over-exploration issue if optimized via policy gradient.

As in the previous setup, we use the training batches of size 1,000 and the learning rate of 0.05 for Adam optimizer, and the evaluation batch of size 10,000 for evaluation.

The simulation results are reported in Figure 3.3 and Table 3.3. As expected, the naïve TS exhibits an extremely poor performance in this setup. At the end of the training process, we observe that all trained algorithms have almost identical performance regardless of the choice of reward metric and baseline in their gradient estimation. During the initial phase of training (i.e., during the first 300 iterations), in contrast, we can observe that the baseline B^{null} performs better than the baseline B^{oracle}. This is in contrast with the heteroscedastic noise example. The intuitive reason is that, in the current setup, the oracle performance does not provide a (nearly) attainable benchmark for an adaptive algorithm.

Figure 3.4 visualizes how the parameterized TS is improved over the course of training. It shows the distribution of pulls that each arm gets, measured at the beginning, middle, and end of the training. At the beginning, since the initial values for meta-parameters are chosen to yield the standard version of TS, it allocates the pulls evenly across the arms, i.e., one pull per one arm in average. As training proceeds, we can observe that the distribution becomes more skewed, i.e., the algorithm effectively ignored some arms as it realizes that it is wasteful to explore all of the arms. The set of ignored arms are randomly determined during the course of policy gradient

optimization: While not reported here, across the choices of reward metric and baseline, the shape of the distribution looks alike but the ordering of arms in the distribution is observed to be different.

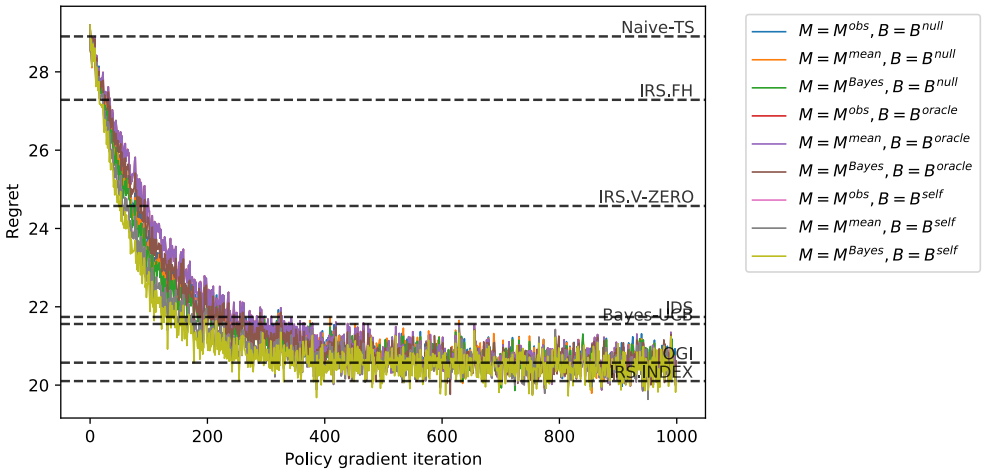

Figure 3.3: Learning curves of parameterized Thompson sampling trained for Gaussian MAB with an excessive number of arms ($K = 20, T = 20$). A curve shows the progress of policy gradient optimization based on a particular choice of reward metric M and baseline B for the gradient estimator $G^{M,B}$, defined in (3.19). The k^{th} data point in each curve reports the average regret on the k^{th} training batch, which contains 1,000 independent instances. The horizontal lines represent the performance of the other algorithms measured with the evaluation batch containing 10,000 instances (also reported in Table 3.3).

Algorithm	Reward metric	Baseline	$\mathrm{Cov}[G^{M,B}]$ ($\times 10^4$)	Regret (s.e.)
Trained TS	M^{obs}	B^{null}	14.99	20.442 (0.125)
	M^{mean}	B^{null}	13.00	20.568 (0.126)
	M^{Bayes}	B^{null}	10.03	20.539 (0.127)
	M^{obs}	B^{oracle}	40.80	20.570 (0.126)
	M^{mean}	B^{oracle}	40.80	20.286 (0.125)
	M^{Bayes}	B^{oracle}	30.73	20.510 (0.125)
	M^{obs}	B^{self}	8.78	20.432 (0.126)
	M^{mean}	B^{self}	8.78	**20.210** (0.125)
	M^{Bayes}	B^{self}	6.07	20.375 (0.126)
NAIVE-TS	–	–	–	28.905 (0.095)
BAYES-UCB	–	–	–	21.561 (0.123)
OGI	–	–	–	20.573 (0.125)
IDS	–	–	–	21.741 (0.119)
IRS.FH	–	–	–	27.286 (0.098)
IRS.V-ZERO	–	–	–	24.575 (0.105)
IRS.INDEX	–	–	–	**20.103** (0.125)

Table 3.3: Performance of the algorithms for Gaussian MAB with an excessive number of arms. Each trained TS uses the meta-parameters found at the end of 1,000 iterations of batched policy gradient ascent (Figure 3.3). The performance is measured in regret, defined in (3.26), and computed via sample average approximation over 10,000 independent instances, and reported with the standard error. The best results are emphasized with bold letters.

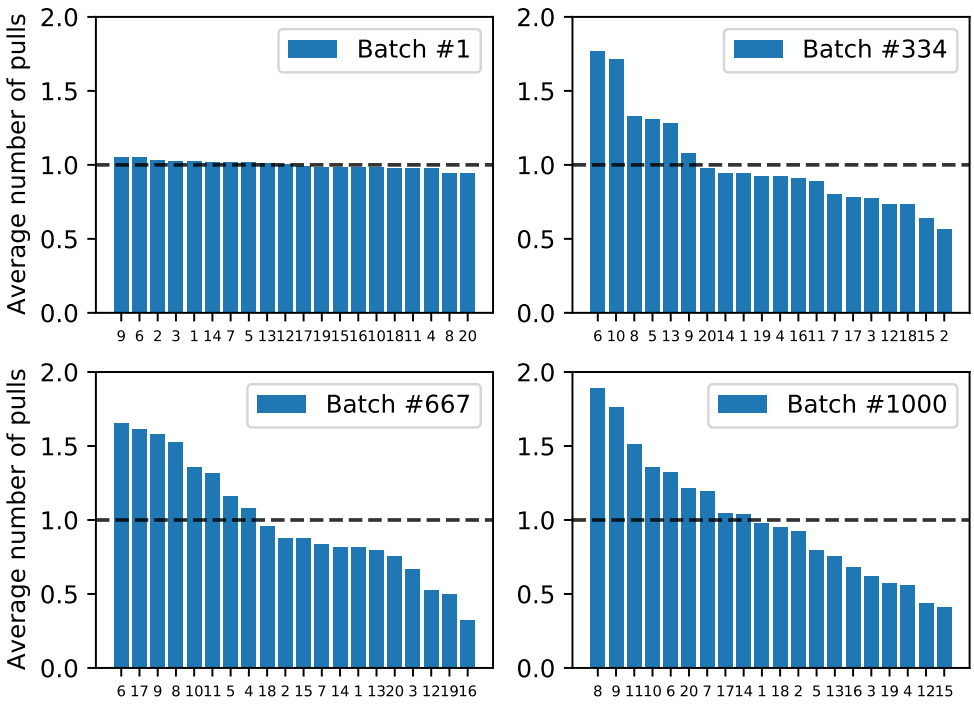

Figure 3.4: Average number of pulls that the parameterized TS conducts on each arm in the k^{th} training batch during the course of policy gradient optimization with reward metric M^{mean} and baseline B^{self}, where $k \in \{1, 334, 667, 1000\}$. Each training batch contains 1,000 independent instances, and the horizontal axes show the arm indices rearranged in the order of the average number of pulls.

Chapter 4: Risk-sensitive Optimal Execution via a Conditional Value-at-Risk Objective

4.1 Introduction

A problem of significant importance for algorithmic traders in modern financial markets is how to allocate their trading efforts over time so as to minimize the cost of trading given a task of liquidation (or acquisition) of a specific amount of an asset. In their seminal paper, [49] framed this liquidation problem as a mean-variance optimization that highlights a tradeoff between the average cost (i.e., the expected implementation shortfall) and the variability of the cost (i.e., the variance of the implementation shortfall). As a result, they have explicitly derived the optimal liquidation schedule that is parameterized with the risk-aversion level of the trader, and their framework and suggested solution have become standards in this area.

However, their analysis is restricted to static strategies: they only considered deterministic schedules under which the algorithm does not adapt to changing market conditions such as the price of the asset. This restriction makes the analysis straightforward because, under a deterministic schedule, the average shortfall only accounts for the transaction cost incurred due to market impact and the variance of the shortfall only accounts for the volatility risk incurred due to random price fluctuation. Under a dynamic strategy, by contrast, its price adaptiveness also contributes to the variability of the outcome, which makes the average term and the risk term entangled in a complicated way.

It has been an important objective to incorporate dynamic strategies into their framework. Some practitioners such as [50] suggest a series of heuristics that are price-adaptive, particularly the one that liquidates more aggressively when the price moves in a favorable direction. [51] observed that this "aggressiveness in-the-money" (AIM) behavior can strictly improve on the optimal deter-

ministic strategy in the mean-variance criterion. In a subsequent paper, [52] develop a dynamic programming technique by which approximate solutions can be obtained, and prove that the optimal (approximate) strategy exhibits AIM, despite the lack of an analytic solution. See also [53] for the continuous-time version of this analysis.

Another stream of work (including this paper) introduces alternative risk criteria other than the mean-variance criterion, so as to formulate the problem into a rather tractable form. For example, [54] formulate the problem as an expected utility maximization problem, and derive a Hamilton–Jacobi–Bellman (HJB) equation that characterizes the optimal adaptive strategy. They find that the optimal strategy is aggressive- or passive-in-the-money, respectively, if and only if the utility function displays increasing or decreasing risk aversion, but an analytic solution is not available. [55] propose an alternative risk criterion that utilizes the time-averaged risk exposure to the price change driven by the geometric Brownian motion (more precisely, the risk term is formulated as the time integration of the position value process, i.e., the product of the position process and the price process), and explicitly solve for the optimal strategy that is shown to exhibit the AIM behavior. [56] investigate the use of the quadratic variation of the position value process as a risk measure, and observe that the classic static solution of [49] is again optimal.[1] [57] introduce a composite dynamic coherent risk measure and derive the optimal solution that is tractable but static. One can also consider an entropic risk measure introduced in [58], but it can be shown that the resulting strategy is also not price-adaptive.

In this paper, we consider the conditional value-at-risk (CVaR; also known as average value-at-risk, tail conditional expectation, or expected shortfall) as a risk measure. In particular, we seek an adaptive liquidation strategy that minimizes the CVaR value of the implementation shortfall in the Almgren–Chriss framework.

The CVaR is a risk measure that quantifies the tail risk. Given a quantile value $q \in (0, 1]$ and a random variable that represents the cost, the CVaR value at level q is defined as the average

[1] This result holds only when the price process is driven by the arithmetic Brownian motion. The authors also consider the geometric Brownian motion under which the optimal solution is shown to be price-adaptive, and they report its AIM behavior through numerical examples.

of the worst q-fraction of the outcomes, i.e., the tail average beyond the q^{th} quantile of the cost distribution. Starting from the pioneering work of [59], it has received much attention for its intuitive definition and for its nice mathematical properties as a coherent risk measure.

From the point of view of the Markov decision process (MDP), the (static) optimization of the CVaR value for a single period can be done efficiently, by utilizing its alternative representation [60]. In the multi-period setting, however, the dynamic optimization of the CVaR value of the total cost is not so trivial. As pointed out by [61] and [62], the optimal action at some point in time may not be completely determined by the current state of the MDP, but may depend on the entire history, and therefore the conventional dynamic programming techniques may not work.

Later studies have adopted the idea of state augmentation to overcome this issue: by introducing an extra state variable, an optimal policy can be sufficiently characterized as a Markov process defined on this augmented state space. Broadly speaking, these studies develop CVaR MDP frameworks using two kinds of state augmentation. The first kind introduces an extra state variable that represents the running cost, and derives the dynamic programming principle from the alternative representation of CVaR, i.e., $\text{CVaR}_q[C] = \min_{c\in\mathbb{R}}\{c + \frac{1}{q}\mathbb{E}[(C-c)^+]\}$ [see 60]. This state augmentation scheme is adopted in e.g., [63], [64], [65], [66], [67]. The second kind of state augmentation introduces an extra state variable that represents the quantile value, and derives the dynamic programming principle from the dual representation of CVaR, i.e., $\text{CVaR}_q[C] = \sup_{\mathbb{Q}:\mathbb{Q}\ll\mathbb{P},\,\frac{d\mathbb{Q}}{d\mathbb{P}}\le q^{-1}} \mathbb{E}_{\mathbb{Q}}[C]$ [see 59]. The work of [68], [69], [70], and [71] belong to this category.

This paper adopts the second kind of state augmentation. More specifically, we consider an augmented state space represented as (X_t, Q_t), where $X_t \in \mathbb{R}$ is the current position size of the trader, and $Q_t \in [0,1]$ is a quantile value that represents the current level of risk aversion. We observe that the dynamic optimization of the CVaR objective can be represented as a (continuous-time) stochastic game between the trader who controls the position process X_t and the adversary who controls the quantile process Q_t. By analyzing the Nash equilibrium of this game, we can identify the minimal CVaR value that the trader can achieve, and specify the trader's optimal policy

and the adversary's optimal policy, which are formulated as time-stationary Markov policies on (X_t, Q_t).

Characterizations of optimal liquidation strategy. Using this approach, we can express the optimal liquidation strategy in an analytic form, and the following observations can be made: (i) the optimal strategy trades in only one direction, i.e., it keeps liquidating until it completes the execution, (ii) it trades more aggressively when the trader is more risk-averse, and (iii) it trades more aggressively when the price moves in a favorable direction, i.e., it is aggressive-in-the-money.

Note that the first property of the optimal strategy is encouraging in practice since the traders (or the clients in the context of a brokerage business) typically do not want to increase their position during the liquidation process. The second property is relatively intuitive in the sense that by liquidating more aggressively the trader can reduce the risk exposure to the price fluctuation more quickly. Most interestingly, the third property can be explained with the second property in our framework. In our game-theoretic interpretation, the quantile process Q_t can be understood as the likelihood that the current sample path ends up with one of worst scenarios that the adversary selects. When a favorable event takes place, it becomes less likely that the sample path is one of the worst scenarios, and therefore the quantile Q_t decreases; i.e., the trader becomes more risk-averse. As a result, the trader is encouraged to liquidate more aggressively by the same reasoning as above.

Contributions. Our contribution is twofold.

First, we derive a tractable solution to the risk-sensitive liquidation problem via a CVaR objective within the Almgren–Chriss framework.[2] As closed-form expressions are available, we can formally characterize the behavior of the optimal liquidation strategy. We can also tractably analyze its performance: compared to the classic static solution of [49], the adaptive strategy can reduce 5–15% of the cost, measured in CVaR; and compared to the volume-weighted average price (VWAP) strategy, it can reduce 15–25% of the cost (see §4.5).

To the best of our knowledge, this is the first work that obtains an analytic solution for the risk-sensitive liquidation problem that is price-adaptive and liquidate-only (i.e., trading in only one

[2]More precisely, we consider a continuous-time and infinite-time version of the Almgren–Chriss framework.

direction). By contrast, the existing work discussed above obtains either a numerical solution [52, 53, 54, 56], a static solution [57], or a strategy that may change its trading direction during the execution [55]. We also believe that the CVaR objective offers a more interpretable quantification of the performance of a trading strategy and a more intuitive control over the risk-aversion level that the trader wants to achieve. However, our results are restricted to the infinite-horizon setting. This may be restrictive in practice, but it is crucial for obtaining a closed-form solution as we can ignore the time dimension when characterizing the value function.

Second, we introduce a novel and technically sound approach to developing the CVaR MDP framework in the continuous-time setting. To sketch our approach briefly, we first introduce a scaled version of CVaR that allows us to avoid the ambiguity of CVaR in the corner case (i.e., when $q = 0$) and inherently induces the concavity of the objective (Proposition 4.2.1). We then utilize the martingale representation theorem, by which we can rewrite the CVaR objective as a maximization problem for an adversary who controls the quantile process Q_t against the decision maker (Theorem 4.3.1), and the problem can be translated into a continuous-time stochastic game between the decision maker and the adversary who are competing over the expected value of the risk-adjusted outcome (Theorem 4.3.2). After this step, we are no longer dealing with the risk-measure, and hence we can safely (time-)decompose the game into the subgames, followed by the CVaR dynamic programming principle (Theorem 4.3.3). This naturally leads to the HJB partial differential equations that characterize the optimal solution. Even though we do exploit the certain structures of the liquidation problem, we believe that our approach provides a promising guideline for developing CVaR MDP frameworks in a broader class of control problems.

We remark that our approach leverages the idea of state augmentation that incorporates the quantile value as an extra state variable. As discussed earlier, this idea has been suggested and utilized in prior work [68, 69, 70, 71], but in discrete-time settings. We clearly take advantage of the continuous-time setting: by exploiting the martingale representation theorem, we can parameterize the adversary's control policy with a real-valued stochastic process that can be tractably optimized. The corresponding optimization in the discrete-time setting is typically much harder

(see the discussion at the end of §4.3.1).

Technical challenges. As we are dealing with a continuous-time and infinite-horizon setting, the analysis involves a considerable amount of technical difficulties. In particular, the distribution of the total cost has an unbounded support in our setting, which makes the analysis much more challenging. We exploit the problem structure to prove the interchangeability of optimization operators, the convergence of solutions, etc. Despite that we make very mild assumptions, we lack a guarantee on the admissibility of the trading strategy that is obtained from the HJB equation. Instead, we prove that this trading strategy can be approximated by admissible ones arbitrarily closely. See Theorem 4.4.3 and the accompanying discussion.

Organization of paper. In §4.2, we introduce the notation and formally describe the model and the problem. In §4.3, we develop the CVaR MDP framework for which we sequentially introduce a martingale representation of the CVaR objective, the game-theoretic representation of the problem, and the Markov policies defined on the augmented state space. In §4.4, we derive the HJB equation, identify its solution, and characterize the optimal liquidation strategy. In §4.5, we compare the optimal adaptive strategy with two deterministic strategies: the optimal deterministic strategy and the optimized VWAP strategy. In §4.6, we provide simulation results that illustrate the optimal strategy. In the appendix, we provide the proofs that are deferred from §4.2–§4.5.

4.2 Problem

We consider a filtered probability space $\big(\Omega, \mathcal{F}, \mathbb{F} = (\mathcal{F}_t)_{t\geq 0}, \mathbb{P}\big)$, where $\mathbb{F}$ is a natural filtration of a Brownian motion $(W_t)_{t\geq 0}$ that satisfies the usual conditions. We denote by $\mathcal{L}^p(\Omega, \mathcal{F}, \mathbb{P})$ (or simply by $\mathcal{L}^p$) the set of $\mathcal{F}$–measurable random variables $X : \Omega \to \mathbb{R}$ such that $\mathbb{E}|X|^p < \infty$. Given a sequence of random variables $(X_n)_{n\in\mathbb{N}}$, it is said that $X_n \xrightarrow{\mathcal{L}^p} X$ if $\lim_{n\to\infty} \mathbb{E}|X_n - X|^p = 0$. The time index set is denoted by $\mathbb{T} \triangleq [0, \infty)$. We also define $\mathcal{P}$ to be the set of progressively measurable stochastic processes in this filtered probability space. We denote $\mathbb{R}_+ \triangleq [0, \infty)$ and $\mathbb{R}_- \triangleq (-\infty, 0]$.

4.2.1 Model

We consider a continuous-time and infinite-horizon version of the setting of [49]. We postulate a trader who wants to liquidate $x \in \mathbb{X}$ units of an asset over an *infinite-time horizon*. Here, x can be negative if the trader wants to acquire the asset. We define $\mathbb{X} \triangleq [-M, M] \subset \mathbb{R}$, where $M \geq 0$ is an arbitrary large number[3] that represents the largest possible position size that the trader is allowed to own.

Liquidation strategy. The trader's liquidation policy is represented with a real-valued continuous-time stochastic process $\pi \triangleq (\pi_t)_{t\geq 0}$, where $\pi_t \in \mathbb{R}$ specifies the *liquidation rate* at time t. Given an initial position size x and a liquidation strategy π, the trader's *position process* $X^{x,\pi} \triangleq (X_t^{x,\pi})_{t\geq 0}$ is determined by

$$X_t^{x,\pi} = x - \int_{s=0}^{t} \pi_s ds; \tag{4.1}$$

i.e., the trader liquidates π_t units of the asset per unit time (or acquires $-\pi_t$ units of the asset per unit time if $\pi_t < 0$). While deferring a formal statement to the end of this subsection, we will restrict our attention to the policies under which the trader's position varies continuously over time (i.e., involves no impulse trades) and vanishes eventually (i.e., converges to zero as t goes to infinity).

Liquidation cost. Following the framework of [49], we define the *cost process* (or loss process) $C^{x,\pi} \triangleq (C_t^{x,\pi})_{t\geq 0}$ as

$$C_t^{x,\pi} \triangleq \int_{s=0}^{t} \frac{1}{2}\eta\pi_s^2 ds - \int_{s=0}^{t} \sigma X_s^{x,\pi} dW_s. \tag{4.2}$$

The first term $\int_{s=0}^{t} \frac{1}{2}\eta\pi_s^2 ds$ represents the (cumulative) transaction cost incurred by the temporary price impact, where the coefficient $\eta > 0$ reflects the illiquidity of the asset. The second term $\int_{s=0}^{t} \sigma X_s^{x,\pi} dW_s$ represents the (cumulative) loss incurred by the random fluctuation of the market price, where $\sigma > 0$ is the volatility of the price process and W_t is the standard Brownian motion. The total cost (i.e., total implementation shortfall) $C_\infty^{x,\pi} \triangleq \lim_{t\to\infty} C_t^{x,\pi}$ is a random variable of interest that we want to minimize via a CVaR objective.

[3]The restriction on the position size is being made to resolve the technical difficulties arising in a convergence analysis. We later show that the choice of M does not play any role in characterizing the optimal strategy; i.e., M does not appear in the optimal solution.

Note that we do not consider permanent price impact in our formulation. Within the Almgren–Chriss framework, it is well known that the contribution of permanent price impact to the implementation shortfall is path-independent; i.e., it does not depend on the liquidation strategy as long as the strategy clears all the positions eventually.[4] Therefore, without loss of generality, we can ignore the presence of permanent impact since it does not make any difference in the trader's decision making. See [49] for details.

Admissible strategies. We now formally define the *set of admissible policies* $\Pi(x)$ as

$$\Pi(x) \triangleq \left\{ \pi : \mathbb{T} \times \Omega \to \mathbb{R} \;\middle|\; \begin{array}{l} \pi \in \mathcal{P}, \\ \mathbb{E}\left[\left(\int_{t=0}^{\infty} \pi_t^2 dt\right)^2\right] < \infty,\ \mathbb{E}\left[\int_{t=0}^{\infty} |X_t^{x,\pi}|^2 dt\right] < \infty, \\ X_t^{x,\pi} \in \mathbb{X},\ \forall t \geq 0 \end{array} \right\}. \tag{4.3}$$

An admissible policy $\pi \in \Pi(x)$ can be dynamic and so it can adjust the trading rate adaptively to the price changes. By this definition, impulse trades are not allowed by the constraint $\mathbb{E}\left[\left(\int_{t=0}^{\infty} \pi_t^2 dt\right)^2\right] < \infty$, and a non-vanishing position is also not allowed by the constraint $\mathbb{E}\left[\int_{t=0}^{\infty} |X_t^{x,\pi}|^2 dt\right] < \infty$. These conditions are to guarantee that the limit $C_\infty^{x,\pi} \triangleq \lim_{t\to\infty} C_t^{x,\pi}$ is an integrable random variable: individual terms in (4.2) converge in $\mathcal{L}^2$ as $t \to \infty$, and therefore $C_t^{x,\pi} \xrightarrow{\mathcal{L}^2} C_\infty^{x,\pi}$ for some $C_\infty^{x,\pi} \in \mathcal{L}^2$.

The last condition $X_t^{x,\pi} \in \mathbb{X} \triangleq [-M, M]$ is seemingly restrictive, but it is merely a technical condition. We allow M to be any arbitrary large number (e.g., $M = 10^{10^{10}}$ units of the asset) so that this constraint will never be restrictive in practice.

4.2.2 Scaled Conditional Value-at-Risk

The *conditional value-at-risk* (CVaR) at a quantile level $q \in (0, 1]$ is a mapping from $\mathcal{L}^1(\Omega, \mathcal{F}, \mathbb{P})$ to $\mathbb{R}$. While there exist several definitions of CVaR in the literature, we consider the one based on

[4] Suppose that liquidating π_t shares of the asset permanently shifts the market price by $-\lambda\pi_t$. Given a liquidation strategy π, the associated transaction cost can be represented as $\int_{t=0}^{\infty} \lambda \pi_t X_t^{x,\pi} dt$, which is always equal to $\frac{1}{2}\lambda x^2$ given that $\lim_{t\to\infty} X_t^{x,\pi} = 0$.

its dual representation as a coherent risk measure [59, 62]: given a random variable $C \in \mathcal{L}^1$,

$$\mathrm{CVaR}_q[C] \triangleq \sup_{Q \in \mathcal{Q}^\dagger(q)} \mathbb{E}[CQ] \quad \text{where} \quad \mathcal{Q}^\dagger(q) \triangleq \left\{ Q \in \mathcal{L}^\infty(\Omega, \mathcal{F}, \mathbb{P}),\ 0 \le Q \le \frac{1}{q},\ \mathbb{E}Q = 1 \right\}, \tag{4.4}$$

i.e., it is defined as a maximization over a random variable Q that has a bounded support $[0, \frac{1}{q}]$ and an expected value of one (or equivalently, a maximization over a probability measure such that it is absolutely continuous with respect to $\mathbb{P}$ and its Radon–Nikodyn derivative with respect to $\mathbb{P}$ is upper bounded by $\frac{1}{q}$ almost surely).

We introduce a scaled version of the CVaR measure, which surprisingly simplifies our analysis.

Definition 4.2.1 (Scaled Conditional Value-at-Risk)**.** *Given a random variable $C \in \mathcal{L}^1(\Omega, \mathcal{F}, \mathbb{P})$ and a quantile $q \in [0, 1]$, the* scaled conditional value-at-risk *(S-CVaR) at level q is*

$$\text{S-CVaR}_q[C] \triangleq \sup_{Q \in \mathcal{Q}(q)} \mathbb{E}[CQ], \tag{4.5}$$

where $\mathcal{Q}(q)$ represents the risk envelope*:*

$$\mathcal{Q}(q) \triangleq \{ Q \in \mathcal{L}^\infty(\Omega, \mathcal{F}, \mathbb{P}) \mid 0 \le Q \le 1,\ \mathbb{E}Q = q \}. \tag{4.6}$$

The S-CVaR measure is obtained by simply scaling the risk envelope of CVaR by q. As a result, $\text{S-CVaR}_q[C]$ is also given by $q \times \mathrm{CVaR}_q[C]$ for $q \neq 0$. Nevertheless, it naturally incorporates the case $q = 0$ into its definition, for which CVaR is not well defined.[5] It further has the following useful properties:

Proposition 4.2.1 (Properties of S-CVaR)**.** *For any random variable $C \in \mathcal{L}^1$, $\text{S-CVaR}_q[C]$ satisfies the following properties:*

(i) $\text{S-CVaR}_q[C] = q \times \mathrm{CVaR}_q[C]$*, for any $q \in (0, 1]$.*

[5] When $q = 0$, $\mathrm{CVaR}_0[C]$ is typically defined as the essential supremum of $C \in \mathcal{L}^1$, which can be infinite if the loss distribution has an unbounded support. We have defined the S-CVaR measure using the dual representation of CVaR so that we can effectively avoid the ambiguity at $q = 0$.

(ii) $\text{S-CVaR}_0[C] = 0$ *and* $\text{S-CVaR}_1[C] = \mathbb{E}C$.

(iii) $\left|\text{S-CVaR}_q[C]\right| \leq \mathbb{E}|C|$ *and* $\text{S-CVaR}_q[C] \geq q\mathbb{E}C$ *for any* $q \in [0, 1]$.

(iv) The mapping $q \mapsto \text{S-CVaR}_q[C]$ *is concave on* $[0, 1]$*, and hence continuous due to (iii).*

(v) Suppose that C *is a continuous random variable whose distribution is atomless. Then,*

$$\text{S-CVaR}_q[C] = \mathbb{E}\left[C\mathbb{I}_{\{C \geq F_C^{-1}(1-q)\}}\right], \tag{4.7}$$

where $F_C^{-1}(\cdot)$ *is the inverse distribution function of* C*, and* $\text{CVaR}_q[C] = \mathbb{E}\left[C \,\middle|\, C \geq F_C^{-1}(1-q)\right]$.

The proof can be found in Appendix B.2. Properties (i)–(iii) provide basic characterizations of S-CVaR. The property (v) provides an interpretation of S-CVaR as a truncated average as opposed to the interpretation of CVaR as a conditional average. We particularly highlight property (iv) that shows the concavity and continuity of the S-CVaR value with respect to q, which is a crucial property that will be exploited in our analysis.

One can interpret the definition (4.5) as a maximization problem for an adversary. This adversary selects a set of scenarios so as to maximize the average cost within the selected scenario, given a constraint that the total measure of the selected scenarios should be q. Informally,[6] the optimized random variable $Q^\star$ is an indicator random variable such that $Q^\star(\omega) = 1$ if the scenario ω is among the worst q-fraction of the scenarios, and $Q^\star(\omega) = 0$ otherwise.

4.2.3 Risk-sensitive execution with a CVaR objective

We now introduce the CVaR risk criterion into the setting described in §4.2.1. In particular, we seek an adaptive strategy that minimizes the CVaR value of the implementation shortfall, given an initial position $x \in \mathbb{X}$ and a target quantile $q \in (0, 1]$. Without loss of generality, we formulate this optimization problem via an S-CVaR objective and define the *value function* $V : \mathbb{X} \times [0, 1] \to \mathbb{R}$

[6] When the loss distribution has an atom at the q^{th} quantile, the extremal random variable $Q^\star(\omega)$ may take a fractional value.

as

$$V(x,q) \triangleq \inf_{\pi\in\Pi(x)} \text{S-CVaR}_q\left[C_\infty^{x,\pi}\right]. \tag{$*$}$$

Note that the above formulation includes the case $q = 0$. By Proposition 4.2.1, the value function $V(x,q)$ is well defined at $q = 0$, and the minimal CVaR value is simply given by $V(x,q)/q$ for any $q \neq 0$. We aim to identify the optimal value function $V(x,q)$ as well as its corresponding optimal liquidation strategy $\pi^\star$.

Recall that the objective $\text{S-CVaR}_q[C_\infty^{x,\pi}]$ concerns the worst q-fraction of outcomes. When $q = 1$, the problem reduces to a risk-neutral liquidation problem. When q takes a smaller value, the problem is equivalent to considering a more risk-averse trader who concerns a smaller fraction of worst cases, being wary of more extreme cases. We anticipate that the trader uses this quantile value $q \in [0,1]$ as an input to our algorithm so as to control the level of risk-aversion that he wants to achieve. In practice, we do not expect the traders to use an extremely small quantile value such as $q = 0.01$ or $q = 0.05$: since they encounter this sort of liquidation task often, possibly on a daily basis, it would be too conservative for them to optimize their performance in the worst 1% or 5% of cases at a cost of sacrificing their performance in the normal 99% or 95% of cases.

4.3 CVaR dynamic programming principle

Using the definition of the S-CVaR measure (4.5), the risk-sensitive optimal execution problem $(*)$ can be formulated as

$$V(x,q) = \inf_{\pi\in\Pi(x)} \sup_{Q\in\mathcal{Q}(q)} \mathbb{E}\left[C_\infty^{x,\pi} Q\right]. \tag{4.8}$$

As discussed in §4.2.2, we can think of an adversary who optimizes a random variable $Q \in \mathcal{Q}(q)$ so as to select the worst q-fraction of sample paths against the trader who employs a liquidation policy $\pi \in \Pi(x)$. In this section, we reformulate the adversary's optimization problem as an optimization over a real-valued continuous-time stochastic process rather than a random variable, and interpret the risk-sensitive optimal control problem as a continuous-time stochastic game between the trader and the adversary. To this end, we develop a continuous-time dynamic programming principle by

exploiting the recursive structure of this game.

4.3.1 Martingale representation of CVaR objective

We consider an arbitrary random variable $C \in \mathcal{L}^1(\Omega, \mathcal{F}, \mathbb{P})$ and derive an alternative representation of S-CVaR$_q[C]$. The results in this subsection are valid not only in the context of the liquidation problem, but also in any filtered probability space generated by a Brownian motion.

We define the *adversary's policy* as a real-valued continuous-time stochastic process $\gamma \triangleq (\gamma_t)_{t\geq 0}$, which determines the *adversary's quantile process* $Q^{q,\gamma} \triangleq (Q_t^{q,\gamma})_{t\geq 0}$:

$$Q_t^{q,\gamma} = q + \int_{s=0}^{t} \gamma_s dW_s, \tag{4.9}$$

where $W = (W_t)_{t\geq 0}$ is the Brownian motion that drives the random price fluctuation. We sometimes call γ_t the *quantile diffusion rate* by analogy to the liquidation rate π_t.

The *set of admissible adversary's policies* is defined as

$$\Gamma(q) \triangleq \left\{\gamma : \mathbb{T}\times\Omega \to \mathbb{R} \;\middle|\; \gamma \in \mathcal{P},\ 0 \leq Q_t^{q,\gamma} \leq 1, \forall t \geq 0\right\}. \tag{4.10}$$

Given an admissible adversary's policy $\gamma \in \Gamma(q)$, its corresponding quantile process $Q^{q,\gamma}$ is a (local) martingale starting at $q \in [0,1]$ whose diffusion term is governed by γ. In particular, the quantile process is required to take values within $[0,1]$ and, as a result, it has the following properties (the proof is provided in Appendix B.3):

Proposition 4.3.1 (Properties of the adversary's quantile process Q)**.** *For any $\gamma \in \Gamma(q)$,*

(i) $(Q_t^{q,\gamma})_{t\geq 0}$ *is a continuous and bounded martingale taking values in* $[0,1]$*, and hence* $\mathbb{E}[Q_\tau^{q,\gamma}] = q$ *for any stopping time* τ*.*

(ii) $Q_\infty^{q,\gamma} \triangleq \lim_{t\to\infty} Q_t^{q,\gamma}$ *exists in* $[0,1]$ *almost surely, and also* $\mathbb{E}[Q_\infty^{q,\gamma}] = q$*.*

(iii) Once $Q_t^{q,\gamma}$ *hits* 0 *or* 1*, it never escapes thereafter.*

By the martingale representation theorem, any random variable Q in the risk envelope $\mathcal{Q}(q)$ can be represented as the limit of a quantile process $Q^{q,\gamma}$ for some $\gamma \in \Gamma(q)$, and vice versa:

Lemma 4.3.1. *For any $q \in [0,1]$, $\mathcal{Q}(q) = \left\{Q^{q,\gamma}_\infty | \gamma \in \Gamma(q)\right\}$.*

Proof. Consider an arbitrary $\tilde{Q} \in \mathcal{Q}(q)$, and Doob martingale $(Q_t)_{t\geq 0}$ generated by $\tilde{Q}$, i.e., $Q_t \triangleq \mathbb{E}[\tilde{Q}|\mathcal{F}_t]$ for each t (we have $\lim_{t\to\infty} Q_t = \tilde{Q}$ and $Q_0 = \mathbb{E}\tilde{Q} = q$). By the martingale representation theorem [72, Thm. 43 in Chap. IV], there exists a predictable process γ such that $Q_t = Q_0 + \int_{s=0}^{t} \gamma_s dW_s$. Since $Q_t = \mathbb{E}[\tilde{Q}|\mathcal{F}_t] \in [0,1]$ for any t, we have an admissible adversary's policy $\gamma \in \Gamma(q)$ and therefore $\tilde{Q} \in \left\{Q^{q,\gamma}_\infty | \gamma \in \Gamma(q)\right\}$ and $\mathcal{Q}(q) \subseteq \left\{Q^{q,\gamma}_\infty | \gamma \in \Gamma(q)\right\}$.

Now consider an arbitrary $\gamma \in \Gamma(q)$ and let $\tilde{Q} \triangleq Q^{q,\gamma}_\infty$. Trivially, $\mathbb{E}\tilde{Q} = Q_0 = q$ and $\tilde{Q} \in [0,1]$. Therefore, $Q^{q,\gamma}_\infty \in \mathcal{Q}(q)$, and hence $\mathcal{Q}(q) \supseteq \left\{Q^{q,\gamma}_\infty | \gamma \in \Gamma(q)\right\}$. □

This alternative representation of the risk envelope $\mathcal{Q}(q)$ immediately leads to the following representation of the S-CVaR value, under which the adversary optimizes over the set of stochastic processes instead of the set of random variables:

Theorem 4.3.1 (Martingale representation of the CVaR objective)**.** *For any given random variable $C \in \mathcal{L}^1(\Omega, \mathcal{F}, \mathbb{P})$ and $q \in [0,1]$, we have*

$$\text{S-CVaR}_q[C] = \sup_{\gamma\in\Gamma(q)} \mathbb{E}\left[CQ^{q,\gamma}_\infty\right]. \tag{4.11}$$

Proof. The result immediately follows from Lemma 4.3.1 and the definition of S-CVaR (4.5). □

Let us consider the optimal martingale $Q^\star$ that solves (4.11), and also recall that the random variable $Q^\star_\infty(\omega)$ indicates whether the sample path ω is among the worst q-fraction of scenarios. As a Doob martingale, the quantile process $Q^\star_t = \mathbb{E}(Q^\star_\infty|\mathcal{F}_t)$ represents its running estimate at time t, i.e., the likelihood that the current sample path will end up with one of the worst scenarios.

Figure 4.1 illustrates a discrete-time analogy of the adversary's quantile process in a two-period setting with $q = \frac{1}{2}$. The adversary will select the worst q-fraction of sample paths (assuming that six sample paths are equally likely to be realized, three out of six sample paths can be selected),

and then each terminal node will be assigned a quantile value 0 or 1. The quantile values at the non-terminal nodes can be sequentially determined in a backward direction by averaging the quantile values of the subsequent nodes, where the quantile value at each node represents, as discussed above, how likely the current sample path ends at one of the terminal nodes selected by the adversary.

We can alternatively interpret this optimization as a sequential decision-making problem that an adversary solves in a forward direction. In Figure 4.1, the root node is assigned a quantile value q, which means that the adversary can select the q-fraction of sample paths realized thereafter. Starting from the root node, the adversary is asked to assign quantile values to the subsequent nodes within a budget of the average of total quantile values he can allocate, where this budget is given by the quantile value at the current node. Then the next state is revealed and the adversary continues to decide how to allocate the quantile values of the next stage until he arrives at a terminal node. The adversary receives a payoff in the amount of the realized cost multiplied by the quantile value allocated to the terminal node.

Note that the probability space generated by a Brownian motion can be well approximated by a binomial model; i.e., each node has two branches that are equally likely to be realized. In the binomial model, the allocation problem that the adversary solves in each period can be represented with a single decision variable: given the current quantile value Q_t, the allocation of next quantile values is of the form $\{(1+\theta_t)Q_t, (1-\theta_t)Q_t\}$, where θ_t is the decision variable. Note that determining the value θ_t is effectively equivalent to determining the diffusion rate of Q_t, which is in fact the adversary's policy γ_t. When there are more than two branches at some node, the allocation problem becomes much harder to solve as it involves more than one decision variable. This illustrates why we can obtain tractable results in the continuous-time setting.

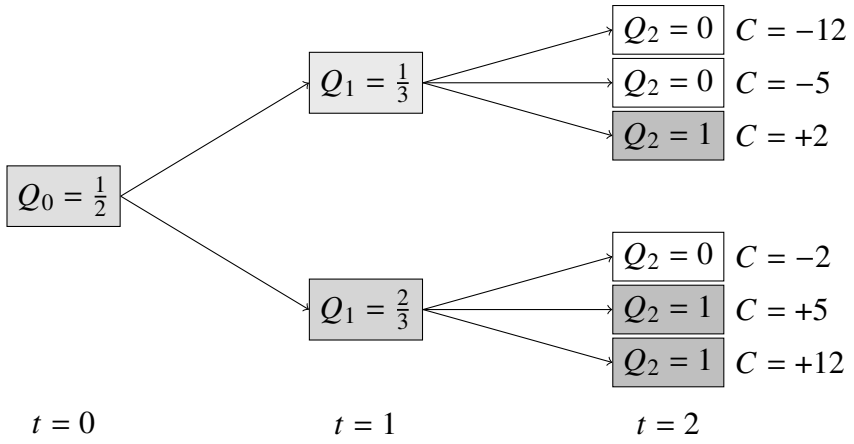

Figure 4.1: An illustration of the discrete-time version of the adversary's martingale in a two-period setting with $q = \frac{1}{2}$. Six sample paths are equally likely to be realized and the label next to each terminal node represents the total cost incurred along each sample path. Interpreting in a backward direction, the adversary selects the q-fraction of terminal nodes (i.e., three out of six nodes at $t = 2$), and the quantile values for the non-terminal nodes are sequentially determined in a backward direction by averaging the quantile values of the subsequent nodes. Interpreting in a forward direction, the adversary starts with a budget of q at the root node, and sequentially allocates the quantile values to the subsequent nodes.

4.3.2 Risk-sensitive liquidation as a continuous-time stochastic game

We now return to the liquidation problem, and define an *outcome function* J as a function of the trader's policy $\pi \in \Pi(x)$ and the adversary's policy $\gamma \in \Gamma(q)$ at each $x \in \mathbb{X}$ and $q \in [0, 1]$:

$$J(\pi, \gamma; x, q) \triangleq \mathbb{E}\left[C_\infty^{x,\pi} Q_\infty^{q,\gamma}\right]. \tag{4.12}$$

By Theorem 4.3.1, the value function $(*)$ can be formulated as

$$V(x, q) = \inf_{\pi \in \Pi(x)} \sup_{\gamma \in \Gamma(q)} J(\pi, \gamma; x, q). \tag{4.13}$$

The following theorem characterizes this value function as an equilibrium outcome of a continuous-time stochastic game between the trader and the adversary.

Theorem 4.3.2 (CVaR optimization as a continuous-time stochastic game)**.** *The value function $V(x, q)$ is the outcome at the Nash equilibrium of the zero-sum game in which the trader and the*

adversary compete over the outcome J:

$$V(x,q) = \inf_{\pi\in\Pi(x)} \sup_{\gamma\in\Gamma(q)} J(\pi,\gamma;x,q) = \sup_{\gamma\in\Gamma(q)} \inf_{\pi\in\Pi(x)} J(\pi,\gamma;x,q). \tag{4.14}$$

Theorem 4.3.2 states that the minimax solution equals the maximin solution. This means that the value function is, as a saddle point, the equilibrium outcome at which each player simultaneously plays the best response against the other player's strategy. This may not always hold true for a general class of risk-sensitive control problems: the convexity of the outcome function with respect to the trader's policy and the convexity of the policy space are required in our proof (Appendix B.3.1) in order to instantiate Sion's minimax theorem [73].

The following remark provides an alternative interpretation of this game based on the Girsanov theorem.

Remark 4.3.1. *The risk-sensitive liquidation problem is equivalent to the risk-neutral liquidation problem in the presence of "price drift" injected by an adversary under a constraint that the likelihood ratio between this altered price process and the original price process cannot exceed $1/q$. In terms of the adversary's martingale $Q^{q,\gamma}$ defined above, the trader solves a risk-neutral liquidation problem where the price process is given by $dP_t = \sigma dW_t + \sigma\frac{\gamma_t}{Q_t^{q,\gamma}}dt$.*

4.3.3 CVaR dynamic programming principle

In this subsection, we develop a continuous-time dynamic programming principle. We first state a proposition that characterizes a temporal structure of the game.

Proposition 4.3.2 (Time decomposition)**.** *Fix $x \in \mathbb{X}$ and $q \in [0,1]$. For any trader's policy $\pi \in \Pi(x)$, adversary's policy $\gamma \in \Gamma(q)$, and a stopping time τ, we have*

$$J(\pi,\gamma;x,q) = \mathbb{E}\left[C_\tau^{x,\pi} Q_\tau^{q,\gamma} + \mathbb{E}\left[\left(C_\infty^{x,\pi} - C_\tau^{x,\pi}\right) Q_\infty^{q,\gamma} \middle| \mathcal{F}_\tau\right]\right]. \tag{4.15}$$

Proof. Observe that $C_\tau^{x,\pi}$ is $\mathcal{F}_\tau$-measurable and $\mathbb{E}[Q_\infty^{q,\gamma}|\mathcal{F}_\tau] = Q_\tau^{q,\gamma}$ since $Q^{q,\gamma}$ is a martingale. Uti-

lizing the tower property, we obtain $J(\pi,\gamma;x,q) = \mathbb{E}\left[C_\infty^{x,\pi}Q_\infty^{q,\gamma}\right] = \mathbb{E}\left[C_\infty^{x,\pi}Q_\infty^{q,\gamma} - (C_\infty^{x,\pi} - C_\tau^{x,\pi})Q_\infty^{q,\gamma}\right] =$ $\mathbb{E}\left[\mathbb{E}\big(C_\tau^{x,\pi}Q_\infty^{q,\gamma}|\mathcal{F}_\tau\big) + \mathbb{E}\big((C_\infty^{x,\pi} - C_\tau^{x,\pi})Q_\infty^{q,\gamma}|\mathcal{F}_\tau\big)\right] = \mathbb{E}\left[C_\tau^{x,\pi}Q_\tau^{q,\gamma} + \mathbb{E}\big((C_\infty^{x,\pi} - C_\tau^{x,\pi})Q_\infty^{q,\gamma}|\mathcal{F}_\tau\big)\right]$. □

Note that $C_\infty^{x,\pi} - C_\tau^{x,\pi}$ represents the cost realized after time τ. Proposition 4.3.2 states that the final outcome can be decomposed into two terms: one term describes the subgame before time τ, and the other term describes the subgame after time τ.

Observe that, in the subgame after time τ, the trader is liquidating $X_\tau^{x,\pi}$ shares and the adversary is selecting a $Q_\tau^{q,\gamma}$-fraction of the future scenarios realized thereafter. This time decomposition naturally leads to the following dynamic programming principle:

Theorem 4.3.3 (CVaR dynamic programming principle)**.** *For any $x \in \mathbb{X}$, $q \in [0,1]$, and a stopping time τ, we have*

$$V(x,q) = \inf_{\pi\in\Pi(x)} \sup_{\gamma\in\Gamma(q)} \mathbb{E}\left[C_\tau^{x,\pi}Q_\tau^{q,\gamma} + V\left(X_\tau^{x,\pi}, Q_\tau^{q,\gamma}\right)\right]. \tag{4.16}$$

Theorem 4.3.3 provides the optimality principle in the form of Bellman's equation: at the equilibrium, the outcome after time τ can be sufficiently described by the subgame equilibrium $V(X_\tau^{x,\pi}, Q_\tau^{q,\gamma})$. The trader is minimizing the (risk-adjusted) cost up to time τ in a consideration of his future state X_τ, while the adversary is simultaneously maximizing the (risk-adjusted) cost up to time τ in a consideration of his future state Q_τ, and the subgame starts at those future states. Like Theorem 4.3.2, Theorem 4.3.3 relies on the saddle-point characterization of the equilibrium; i.e., it does not matter which player commits his policy first in the subgame. The formal proof can be found in Appendix B.3.3.

4.3.4 (X,Q)-Markov policies

Theorem 4.3.3 implies that the *augmented state space* (X_t, Q_t) is sufficient to describe the remaining subgame at time t. Therefore, if a policy is reasonable, its action at time t (π_t or γ_t) should be determined by the current position size X_t and the current quantile value Q_t. To formalize this idea, we introduce time-stationary Markov policies running on this augmented state space:

Definition 4.3.1 ((X, Q)-Markov policies). *We say that a trader's policy π is an (X, Q)-Markov policy coupled with γ if*

$$\pi_t(\omega) = f\left(X_t^{x,\pi}(\omega), Q_t^{q,\gamma}(\omega)\right), \quad \forall t, \omega \tag{4.17}$$

for some measurable function $f : \mathbb{R} \times [0, 1] \to \mathbb{R}$.

Similarly, an adversary's policy γ is an (X, Q)-Markov policy coupled with π if

$$\gamma_t(\omega) = g\left(X_t^{x,\pi}(\omega), Q_t^{q,\gamma}(\omega)\right) \tag{4.18}$$

for some measurable function $g : \mathbb{R} \times [0, 1] \to \mathbb{R}$.

A policy pair (π, γ) is a mutually coupled (X, Q)-Markov policy pair if both (4.17) *and* (4.18) *hold.*

An (X, Q)-Markov policy is characterized by a function defined on the augmented state space. The function f or g specifies the liquidation rate or the quantile diffusion rate when the current position size is x and the current quantile level is q. Recall that we have defined a policy, π or γ, as a continuous-time stochastic process adapted to the Brownian motion, i.e., as a (progressively measurable) mapping $\pi : \mathbb{T} \times \Omega \to \mathbb{R}$. Strictly speaking, the function f or g does not completely determine one player's policy unless the other player's policy is specified. To avoid this ambiguity, when we describe an (X, Q)-Markov policy, we specify the other player's policy that is coupled with it.

To better understand, consider the policies π and γ that are mutually coupled (X, Q)-Markov policies induced by functions f and g. Under π and γ, the system is completely described by the coupled processes $(X_t, Q_t)_{t\geq 0}$ on the augmented state space, whose dynamics are given by the following stochastic differential equations:

$$dX_t = -f\left(X_t, Q_t\right) dt, \quad dQ_t = g\left(X_t, Q_t\right) dW_t, \tag{4.19}$$

with the initial states $X_0 = x$ and $Q_0 = q$. Even if γ is not an (X, Q)-Markov policy, we can still

consider an (X, Q)-Markov policy π that is induced by f and coupled with γ, and then the position process will be given by $dX_t = -f\left(X_t, Q_t^{q,\gamma}\right) dt$.

Note also that the admissibility of an (X, Q)-Markov policy is not always guaranteed: it may fail to satisfy the admissible conditions given in (4.3) or (4.10), depending on the generating function f or g as well as the other player's policy coupled with it. See the discussions before and after Theorem 4.4.3.

4.4 Optimal solution

In this section, we utilize the CVaR dynamic programming principle (Theorem 4.3.3) to derive a Hamilton–Jacobi–Bellman (HJB) equation for the risk-sensitive optimal execution problem, and identify the functional form of the value function and optimal policies by solving this HJB equation. For all propositions/theorems stated in this section, we defer their proofs to Appendix B.4.

4.4.1 Minimal CVaR cost

We first state Dynkin's formula that we can apply to the right-hand side of (4.16) in Theorem 4.3.3 so as to represent it as a time-integration.

Proposition 4.4.1 (Dynkin's formula)**.** *Consider a function $\widehat{V} : \mathbb{R} \times [0,1] \to \mathbb{R}$ such that $\widehat{V} \in C^{1,2}(\mathbb{R} \times (0,1))$. For any $\pi \in \Pi(x)$, $\gamma \in \Gamma(q)$, and a stopping time τ such that $Q_\tau \in (0,1)$ almost surely, we have*

$$\mathbb{E}\left[C_\tau^{x,\pi} Q_\tau^{q,\gamma} + \widehat{V}\left(X_\tau^{x,\pi}, Q_\tau^{q,\gamma}\right)\right] = \widehat{V}(x,q) \tag{4.20}$$

$$+ \mathbb{E}\left[\int_{t=0}^{\tau} \left\{\frac{\eta}{2} Q_t^{q,\gamma} \pi_t^2 - \widehat{V}_x\left(X_t^{x,\pi}, Q_t^{q,\gamma}\right) \pi_t\right\} dt\right] \tag{4.21}$$

$$+ \mathbb{E}\left[\int_{t=0}^{\tau} \left\{\frac{1}{2} \widehat{V}_{qq}\left(X_t^{x,\pi}, Q_t^{q,\gamma}\right) \gamma_t^2 - \sigma X_t^{x,\pi} \gamma_t\right\} dt\right], \tag{4.22}$$

where $\widehat{V}_x(x,q) \triangleq \frac{\partial}{\partial x}\widehat{V}(x,q)$ and $\widehat{V}_{qq}(x,q) \triangleq \frac{\partial^2}{\partial q^2}\widehat{V}(x,q)$.

For the sake of argument, suppose that the value function V is twice differentiable so that it can be plugged into Proposition 4.4.1 in the place of $\widehat{V}$. When considering an infinitesimal time interval (i.e., $\tau = dt$), we have

$$V(x,q) \overset{\text{Thm 4.3.3}}{=} \inf_{\pi\in\Pi(x)} \sup_{\gamma\in\Gamma(q)} \mathbb{E}\left[C_\tau^{x,\pi} Q_\tau^{q,\gamma} + V\left(X_\tau^{x,\pi}, Q_\tau^{q,\gamma}\right)\right] \tag{4.23}$$

$$\overset{\text{Prop 4.4.1}}{=} \inf_{\pi\in\Pi(x)} \sup_{\gamma\in\Gamma(q)} \left\{V(x,q) + \left(\frac{\eta}{2} q\pi_0^2 - V_x(x,q)\,\pi_0\right) dt + \left(\frac{1}{2} V_{qq}(x,q)\,\gamma_0^2 - \sigma x \gamma_0\right) dt\right\} \tag{4.24}$$

$$= V(x,q) + \inf_{\pi\in\Pi(x)} \left\{\frac{\eta}{2} q\pi_0^2 - V_x(x,q)\,\pi_0\right\} dt + \sup_{\gamma\in\Gamma(q)} \left\{\frac{1}{2} V_{qq}(x,q)\,\gamma_0^2 - \sigma x\gamma_0\right\} dt. \tag{4.25}$$

Observe that the terms associated with the trader's policy π and the terms associated with the adversary's policy γ can be separated. We can naturally infer that the value function V has to satisfy the following HJB equation:

$$\min_{v\in\mathbb{R}} \left\{\frac{\eta}{2} qv^2 - V_x(x,q)\,v\right\} + \max_{w\in\mathbb{R}} \left\{\frac{1}{2} V_{qq}(x,q)\,w^2 - \sigma x w\right\} = 0. \tag{4.26}$$

In the following theorem, we make this argument more formally and provide sufficient characterizations of the value function.

Theorem 4.4.1 (Verification theorem)**.** *Consider a function $V^\star : \mathbb{R}\times[0,1] \to \mathbb{R}_+$ satisfying*

(i) $V^\star \in C^{1,2}(\mathbb{R}\times(0,1))$, and, for any $x\in\mathbb{R}$ and $q\in(0,1)$,

$$\min_{v\in\mathbb{R}} \left\{\frac{\eta}{2} qv^2 - V_x^\star(x,q)\,v\right\} + \max_{w\in\mathbb{R}} \left\{\frac{1}{2} V_{qq}^\star(x,q)\,w^2 - \sigma x w\right\} = 0, \tag{4.27}$$

where $V_x^\star \triangleq \frac{\partial V^\star}{\partial x}$ and $V_{qq}^\star \triangleq \frac{\partial^2 V^\star}{\partial q^2}$.

(ii) $V^\star(0,q) = 0$ for all $q\in[0,1]$, and $V^\star(x,0) = V^\star(x,1) = 0$ for all $x\in\mathbb{R}$.

(iii) $V^\star(x,q) = V^\star(-x,q)$ for all $x\in\mathbb{R}$ and $q\in[0,1]$, and $V^\star(x,q)$ is increasing in x on $\mathbb{R}_+$.

(iv) $\frac{V^\star_x(x,q)}{q}$ *is continuous on* $\mathbb{R}_+ \times (0,1)$*, increasing in* x *on* $\mathbb{R}$ *for each* $q \in (0,1)$*, and decreasing in* q *on* $(0,1)$ *for each* $x \in \mathbb{R}$*.*

(v) $\frac{x}{V^\star_{qq}(x,q)}$ *is continuous on* $\mathbb{R}_+ \times (0,1)$*, decreasing in* x *on* $\mathbb{R}_+$*.*

Then, $V(x,q) = V^\star(x,q)$ *for all* $x \in \mathbb{X}$ *and* $q \in [0,1]$*.*

In Theorem 4.4.1, conditions (i) and (ii) specify the HJB equation and the boundary conditions that the value function has to satisfy. In fact, the value function V can be uniquely determined by these two conditions. However, the other conditions, (iii)–(v), are also necessary to show that this value is indeed achievable within our definition of the admissible policies, $\Pi(x)$ and $\Gamma(q)$. More specifically, condition (iii) asserts the symmetry and the monotonicity of the value function with respect to the position size x, and conditions (iv) and (v) assert certain behaviors of the optimal policies that are implied from the HJB equation (e.g., the optimal liquidation strategy should trade more aggressively when liquidating a larger quantity). While these properties of the value function are what naturally follow from the problem structure, they serve as regularity conditions in our proof to resolve technical issues arising in the convergence analysis.

Observe that the optimization terms in the HJB equation (4.27) are separated and each of them is a trivial quadratic optimization problem. By solving these optimizations explicitly, the HJB equation can be translated into the following partial differential equation:

$$V_x^2(x,q) \times V_{qq}(x,q) = -\sigma^2 \eta \times x^2 \times q. \tag{4.28}$$

It turns out that this differential equation with the boundary condition (ii) admits a separable solution. The value function $V(x,q)$ can be represented as a product of a function of x and a function of q, as identified in the following theorem.

Theorem 4.4.2 (Value function)**.** *Consider a function* $V^\star : \mathbb{R} \times [0,1] \to \mathbb{R}_+$ *defined as*

$$V^\star(x,q) \triangleq (3/4)^{\frac{2}{3}} \times \sigma^{\frac{2}{3}} \eta^{\frac{1}{3}} \times |x|^{\frac{4}{3}} \times \varphi(q), \tag{4.29}$$

where $\varphi : [0,1] \to \mathbb{R}_+$ *is the solution in* $C((0,1))$ *to the following differential equation:*

$$\varphi^2(q) \times \varphi''(q) = -q, \ \forall q \in (0,1), \quad \textit{and} \quad \varphi(0) = \varphi(1) = 0. \tag{4.30}$$

Then, $V^\star$ *satisfies the conditions of Theorem 4.4.1, and hence* $V(x,q) = V^\star(x,q)$ *for all* $x \in \mathbb{R}$ *and* $q \in [0,1]$.

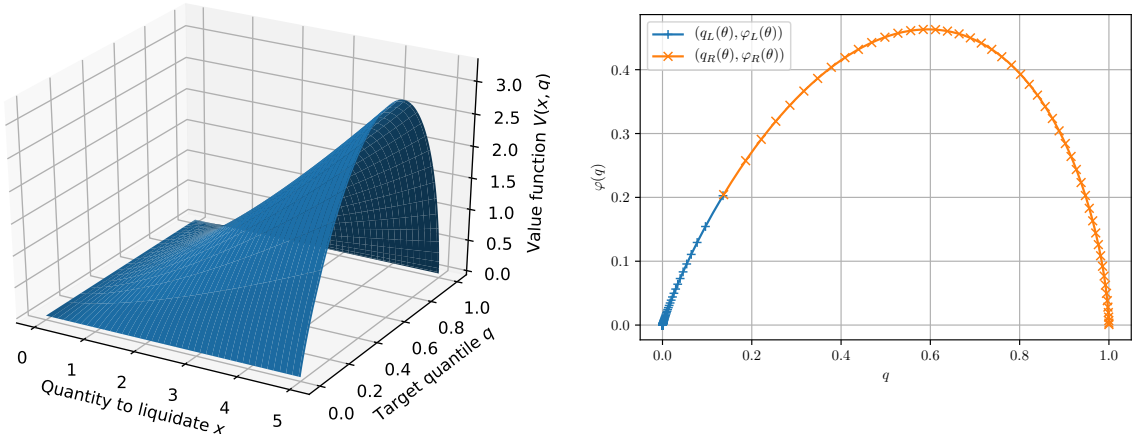

Figure 4.2: (Left) Value function $V(x,q)$ that represents the minimal S-CVaR loss (i.e., $\text{S-CVaR}_q[C_\infty]$) at a quantile level q when liquidating x units of an asset given that $\sigma = \eta = 1$. The closed-form expression is provided in (4.29). (Right) Function $\varphi(q)$ that is derived in Proposition 4.4.2. The curve $\{(q, \varphi(q))\}_{q\in[0,1]}$ is represented in parametric form, in which the left part of the curve is represented as $\{(q_L(\theta), \varphi_L(\theta))\}_{\theta\in[0,\infty]}$, and the right part of the curve is represented as $\{(q_R(\theta), \varphi_R(\theta))\}_{\theta\in[0,\bar{\theta}]}$.

The differential equation of form (4.30) is known as the *Emden–Fowler equation* [74, p. 2.3.27], and its solution can be expressed in parametric form as follows:

Proposition 4.4.2 (Parametric representation of $\varphi(q)$)**.** *The function* $\varphi(q)$ *can be represented in a parametric form that admits a closed-form expression. Define*

$$Z_L(\theta) \triangleq -\frac{2}{\pi} K_{1/3}(\theta), \quad Z_R(\theta) \triangleq \sqrt{3} J_{1/3}(\theta) - Y_{1/3}(\theta) \tag{4.31}$$

where J *and* Y *are the first and second kinds of Bessel functions, and* K *is the second kind of*

modified Bessel function. Further define

$$\bar{\theta} \triangleq \inf\{\theta > 0 : Z_R(\theta) = 0\} \approx 2.3834, \quad a \triangleq \frac{1}{\bar{\theta}^{\frac{4}{3}} \left(Z_R'(\bar{\theta})\right)^2} \approx 0.2910, \quad b \triangleq a\,(9/2)^{\frac{1}{3}} \approx 0.1763. \tag{4.32}$$

Then, the curve $\{(q, \varphi(q))\}_{q\in[0,1]}$ *is parameterized as*

$$\{(q, \varphi(q))\}_{q\in[0,1]} = \{(q_L(\theta), \varphi_L(\theta))\}_{\theta\in[0,\infty]} \bigcup \{(q_R(\theta), \varphi_R(\theta))\}_{\theta\in[0,\bar{\theta}]}, \tag{4.33}$$

where[7]

$$q_L(\theta) \triangleq a\theta^{-\frac{2}{3}} \left[\left(\theta Z_L'(\theta) + \frac{1}{3} Z_L(\theta)\right)^2 - \theta^2 Z_L^2(\theta)\right], \quad \varphi_L(\theta) \triangleq b\theta^{\frac{2}{3}} Z_L^2(\theta), \tag{4.34}$$

and

$$q_R(\theta) \triangleq a\theta^{-\frac{2}{3}} \left[\left(\theta Z_R'(\theta) + \frac{1}{3} Z_R(\theta)\right)^2 + \theta^2 Z_R^2(\theta)\right], \quad \varphi_R(\theta) \triangleq b\theta^{\frac{2}{3}} Z_R^2(\theta). \tag{4.35}$$

4.4.2 Optimal adaptive liquidation strategy

Let $f^\star(x, q)$ and $g^\star(x, q)$ be, respectively, the minimizer and the maximizer of the optimization terms in the HJB equation (4.27), i.e.,

$$f^\star(x, q) \triangleq \underset{v\in\mathbb{R}}{\operatorname{argmin}} \left\{\frac{\eta}{2} q v^2 - V_x(x, q)\, v\right\} = \frac{V_x(x, q)}{\eta q} = (3/4)^{-\frac{1}{3}} \times \sigma^{\frac{2}{3}} \eta^{-\frac{2}{3}} \times x^{\frac{1}{3}} \times \frac{\varphi(q)}{q}, \tag{4.36}$$

$$g^\star(x, q) \triangleq \underset{w\in\mathbb{R}}{\operatorname{argmax}} \left\{\frac{1}{2} V_{qq}(x, q)\, w^2 - \sigma x w\right\} = \frac{\sigma x}{V_{qq}(x, q)} = -(3/4)^{-\frac{2}{3}} \times \sigma^{\frac{1}{3}} \eta^{-\frac{1}{3}} \times x^{-\frac{1}{3}} \times \frac{\varphi^2(q)}{q}, \tag{4.37}$$

where we define $x^{\frac{1}{3}} = -|x|^{\frac{1}{3}}$ for $x < 0$. The function $f^\star(x, q)$ specifies the trader's optimal liquidation rate when the current position size is x and the current quantile level is q, and similarly, the function $g^\star(x, q)$ specifies the adversary's optimal quantile diffusion rate in that situation.

[7] The values of $Z_L(\theta)$ and $Z_R(\theta)$ are not defined when $\theta = 0$. However, the limit points (e.g., $\lim_{\theta\searrow 0}(q_L(\theta), \varphi_L(\theta))$, $\lim_{\theta\nearrow\infty}(q_L(\theta), \varphi_L(\theta))$) do exist, and our parametric representation includes those limit points.

We can naturally postulate mutually coupled (X, Q)-Markov policies $\pi^\star$ and $\gamma^\star$ induced by these functions $f^\star$ and $g^\star$, which are characterizing the equilibrium of the stochastic game.[8] Under policies $\pi^\star$ and $\gamma^\star$, the system is described by the following stochastic differential equations:

$$dX_t = -f^\star(X_t, Q_t)\, dt, \quad dQ_t = g^\star(X_t, Q_t)\, dW_t, \tag{4.38}$$

with $X_0 = x$ and $Q_0 = q$.

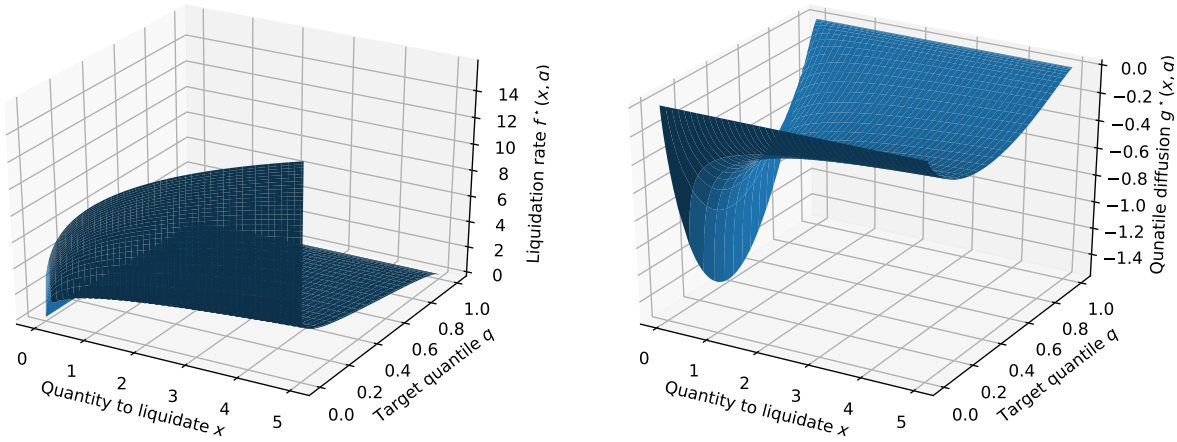

Figure 4.3: Optimal trading rate function $f^\star(x, q)$ (left) and optimal quantile diffusion rate function $g^\star(x, q)$ (right) given that $\sigma = \eta = 1$. The closed-form expressions are provided in (4.36) and (4.37).

However, we cannot directly show that the policies $\pi^\star$ and $\gamma^\star$ satisfy the admissibility conditions introduced in (4.3) and (4.10), because the functions $f^\star$ and $g^\star$ exhibit extreme behaviors near the boundaries, as shown in Figure 4.3. For example, if $Q_t \searrow 0$, the liquidation rate π_t may increase unboundedly since $\lim_{q \searrow 0} f^\star(x, q) = \infty$, and thus may violate the condition $\mathbb{E}\left[\left(\int_{t=0}^{\infty} \pi_t^2 dt\right)^2\right] < \infty$. If $Q_t \nearrow 1$, on the other hand, the liquidation rate π_t may vanish since $\lim_{q \nearrow 1} f^\star(x, q) = 0$, and thus may violate the condition $\mathbb{E}\left[\int_{t=0}^{\infty} |X_t^{x,\pi}|^2 dt\right] < \infty$.

[8] This does not mean that the policy $\pi^\star$ is optimal against any adversary's policy γ, nor an (X, Q)-Markov policy induced by $f^\star$ and coupled with γ is the best response against γ. It merely means that $\pi^\star$ is the best response against $\gamma^\star$ only, and vice versa. In order to obtain a best response against an arbitrary adversary's policy $\gamma \in \Gamma(q)$, we may need to characterize the best possible performance against γ, e.g., $V(x, q; \gamma) \triangleq \inf_{\pi \in \Pi(x)} J(\pi, \gamma; x, q)$, and derive and solve the HJB equation associated with it. Nevertheless, the policy $\pi^\star$ is the optimal liquidation strategy that minimizes the CVaR value of implementation shortfall.

We instead prove that the optimal value function can be achieved "asymptotically" by a sequence of admissible policies that approximate $\pi^\star$ and $\gamma^\star$.

Theorem 4.4.3 (Policy optimality)**.** *There exists a sequence of function pairs* $(f^{(n)}, g^{(n)})_{n\in\mathbb{N}}$ *such that*

$$\lim_{n\to\infty} f^{(n)}(x,q) = f^\star(x,q), \quad \lim_{n\to\infty} g^{(n)}(x,q) = g^\star(x,q), \quad \forall (x,q) \in \mathbb{R}\setminus\{0\}\times(0,1], \tag{4.39}$$

and it satisfies the following properties for any $x \in \mathbb{X}$ *and* $q \in [0,1]$*:*

(a) For any given $\gamma \in \Gamma(q)$*, let* $\pi^{(n),\gamma}$ *be an* (X,Q)*-Markov trader's policy induced by* $f^{(n)}$ *and coupled with* γ*. Then,* $\pi^{(n),\gamma}$ *is admissible and*

$$\limsup_{n\to\infty} J(\pi^{(n),\gamma}, \gamma; x, q) \le V(x,q), \quad \forall \gamma \in \Gamma(q). \tag{4.40}$$

(b) For any given $\pi \in \Pi(x)$*, let* $\gamma^{(n),\pi}$ *be an* (X,Q)*-Markov adversary's policy induced by* $g^{(n)}$ *and coupled with* π*. Then,* $\gamma^{(n),\pi}$ *is admissible and*

$$\liminf_{n\to\infty} J(\pi, \gamma^{(n),\pi}; x, q) \ge V(x,q), \quad \forall \pi \in \Pi(x). \tag{4.41}$$

(c) Let $(\pi^{(n)}, \gamma^{(n)})$ *be a mutually coupled* (X,Q)*-Markov policy pair induced by* $(f^{(n)}, g^{(n)})$*. Then,* $\pi^{(n)}$ *and* $\gamma^{(n)}$ *are admissible and*

$$\lim_{n\to\infty} J(\pi^{(n)}, \gamma^{(n)}; x, q) = V(x,q). \tag{4.42}$$

Theorem 4.4.3 shows that we can construct a sequence of functions $(f^{(n)}, g^{(n)})_{n\in\mathbb{N}}$ that converges to $(f^\star, g^\star)$ pointwise except at the boundaries, and further induces (X,Q)-Markov policies that are admissible and asymptotically optimal. More precisely, against any adversary's policy γ, the sequence of functions $(f^{(n)})_{n\in\mathbb{N}}$ induces a sequence of admissible policies $(\pi^{(n),\gamma})_{n\in\mathbb{N}}$ for the

trader, and in the limit, the trader achieves an outcome that is no worse than the equilibrium outcome. And vice versa, against any trader's policy π, the sequence of functions $(g^{(n)})_{n\in\mathbb{N}}$ induces a sequence of admissible policies $(\gamma^{(n),\pi})_{n\in\mathbb{N}}$ for the adversary, and in the limit, the adversary achieves an outcome that is no worse than the equilibrium outcome. As a combination, the sequence of function pairs $(f^{(n)}, g^{(n)})_{n\in\mathbb{N}}$ induces a sequence of mutually admissible policy pairs $(\pi^{(n)}, \gamma^{(n)})_{n\in\mathbb{N}}$ that yields the equilibrium outcome asymptotically.

The construction of such a sequence of function pairs $(f^{(n)}, g^{(n)})_{n\in\mathbb{N}}$ is straightforward. We consider a vanishing subset of the augmented state space that contains the boundaries, i.e., $\{(x,q)||x| \leq \frac{1}{n}, q \leq \frac{1}{n} \text{ or } q \geq 1-\frac{1}{n}\}$, and obtain $f^{(n)}$ and $g^{(n)}$ by suppressing the extreme behaviors of $f^\star$ and $g^\star$ arising in this subset. Roughly speaking, the liquidation strategy induced by $(f^{(n)}, g^{(n)})$ mimics the optimal strategy until it clears almost all positions (i.e., $X_t \leq \frac{1}{n}$) or it becomes almost risk-neutral (i.e., $Q_t \geq 1-\frac{1}{n}$) or extremely risk-averse (i.e., $Q_t \leq \frac{1}{n}$), and then liquidates according to a deterministic schedule thereafter. We can show that the gap between the outcome of this approximated strategy and the theoretical equilibrium outcome is diminishing as n goes to infinity. We refer the readers to Appendix B.4 for the details.

Aggressiveness-in-the-money. Despite that the admissibility of the optimal liquidation policy $\pi^\star$ is not guaranteed, we can still characterize its behaviors by inspecting the stochastic differential equations (4.38). Without loss of generality, let us consider the task of liquidation (i.e., $x \geq 0$).

First, we observe that the optimal policy liquidates only (i.e., $f^\star(x,q) \geq 0$), until it completes the execution[9] (i.e., $f^\star(0,q) = 0$). Note that we have not imposed any constraint on the trading direction. This formally shows that winding back the position during the liquidation process will never be helpful in reducing the CVaR loss.

Second, when the trader becomes more risk-averse (i.e., $q \searrow 0$), the optimal policy trades more aggressively (i.e., $f^\star(x,q) \nearrow \infty$). The opposite also holds true. This is because, by liquidating the position more quickly, it can reduce the risk exposure to the changes in price more effectively.

[9] We are not sure if the completion time is almost surely finite even though the position will be vanishing eventually (i.e., $X_t \searrow 0$). In particular, when $Q_t \approx 1$, the optimal policy trades very slowly ($\pi_t \approx 0$) and the position process may never hit zero. We believe that the completion time is finite with a probability of at least $1-q$.

Even though it will be more costly in terms of market impact, it can make sure that the transactions will be made at a certain level, which is more beneficial to a risk-averse trader than a risk-neutral trader. We can also understand this behavior based on the alternative interpretation of the problem discussed in Remark 4.3.1: when the risk-averse liquidation problem is cast as a risk-neutral execution problem that involves an adverse price drift, being more risk-averse is equivalent to facing a more adverse price drift, which encourages the risk-neutral trader to trade more aggressively.

Most interestingly, we observe that when the price moves in a favorable direction toward in-the-money (i.e., $dW_t > 0$), the policy becomes more risk-averse (i.e., $dQ_t < 0$ since $g^\star(x, q) < 0$) and hence it trades more aggressively. This formally characterizes aggressiveness-in-the-money, which has been observed by [51, 52, 55, 56]. Intuitively, if the trader has made some "free" money due to the price change, he would have an additional incentive to complete the liquidation early so as to secure his current profit, and thus he would be willing to pay an additional deterministic cost for aggressive execution.

This behavior can also be understood in the context of a more general risk-sensitive optimal control problem. Recall that the optimal adversary's martingale $Q_t^{q,\gamma^\star}$ represents the likelihood that the current sample path leads to one of the worst q-fraction of outcomes. When something favorable happens, it becomes less likely that the sample path is in the worst q^{th} quantile, and therefore Q_t decreases. This means that the trader will need to pay attention to a smaller fraction of adverse scenarios; i.e., he will become more risk-averse.

A threshold behavior. While we do not have a formal characterization here, we observe that the optimal strategy exhibits some threshold behavior, particularly near the end of the liquidation. We observe that the policy trades aggressively when the cumulative cost C_t is below some threshold, and it trades passively when the cumulative cost C_t is above the threshold. Near the end of the liquidation, such a threshold corresponds to $\text{VaR}_q[C_\infty]$ (the value-at-risk, i.e., the q^{th} quantile of the loss distribution), and the liquidation rate sharply changes around $\text{VaR}_q[C_\infty]$. This behavior is related to an alternative representation of CVaR [60]: $\text{CVaR}_q[C_\infty] = \max_{c\in\mathbb{R}}\{c + \frac{1}{q}\mathbb{E}\left[(C_\infty - c)^+\right]\}$, where the maximizer $c^\star$ is in fact $\text{VaR}_q[C_\infty]$, and $\text{CVaR}_q[C_\infty]$ only concerns the cases where

$C_\infty > c^\star$.

To better illustrate, suppose that the trader is currently left with a small amount of position to liquidate. If $C_t < c^\star$, the trader is willing to pay a large transaction cost (up to $c^\star - C_t$) to complete the execution as soon as possible, thereby making sure that the total loss C_∞ won't exceed the threshold $c^\star$. If $C_t > c^\star$, the trader may believe that the total loss C_∞ will inevitably exceed the threshold $c^\star$, and then tries to minimize the expected future cost by slowing down the liquidation. One can make a connection with aggressiveness-in-the-money, since having $C_t < c^\star$ implies that the price has moved in a favorable direction.

We believe that the threshold value is a function of remaining position size such that it increases as the position size decreases and converges to $\mathrm{VaR}_q[C_\infty]$ as the position vanishes. Moreover, the change in the aggressiveness around the threshold also depends on the remaining position size. To formalize this behavior, one may adopt an alternative formulation of the problem with an extra state variable representing the cost realized so far (i.e., a Markov policy defined on the augmented state space (X_t, C_t)), which is in fact the approach suggested by [63] for a general class of control problems with a CVaR objective. This might be a topic of future research.

4.5 Cost analysis: adaptive vs. deterministic strategy

In this section, we provide a comparison between the optimal adaptive strategy derived in §4.4 and the (optimized) deterministic schedules under which the liquidation is executed according to a deterministic schedule committed at the beginning of the liquidation process.

4.5.1 Optimized deterministic schedules

First observe that any deterministic schedule will yield a normally distributed implementation shortfall; i.e., given a deterministic schedule π, the total implementation shortfall $C_\infty^{x,\pi} = \int_0^\infty \frac{\eta}{2}\pi_t^2 dt - \int_0^\infty \sigma X_t dW_t$ follows a normal distribution with a mean $\mathbb{E}[C_\infty^{x,\pi}] = \int_0^\infty \frac{\eta}{2}\pi_t^2 dt$ and a variance $\mathrm{Var}[C_\infty^{x,\pi}] = \int_0^\infty \sigma^2 X_t^2 dt$. To investigate the performance of deterministic schedules, therefore, it suffices to formulate the CVaR value of a normal distribution.

Given a random variable Z distributed with $\mathcal{N}(0, 1^2)$, we let $\kappa(q)$ be its CVaR value which admits the following expression:

$$\kappa(q) \triangleq \mathrm{CVaR}_q[Z] = \phi(\Phi^{-1}(1-q)), \tag{4.43}$$

where $\phi(\cdot)$ and $\Phi(\cdot)$ are the p.d.f. and the c.d.f. of a standard normal distribution, respectively. Consequently, we have that $\mathrm{CVaR}_q[C] = \mu + \sigma\kappa(q)$ for any random variable C distributed with $\mathcal{N}(\mu, \sigma^2)$, and therefore, $\mathrm{CVaR}_q[C_\infty^{x,\pi}] = \int_0^\infty \frac{\eta}{2}\pi_t^2 dt + \kappa(q)\sqrt{\int_0^\infty \sigma^2 X_t^2 dt}$ for any deterministic schedule π.

We first focus on the set of all deterministic schedules and find the optimal one that minimizes the CVaR cost. The next proposition shows that such an optimal schedule has the form of an exponential schedule under which the trader's position decays exponentially over time (i.e., the liquidation rate is proportional to the current position size).

Proposition 4.5.1 (Optimized deterministic schedule)**.** *Given an initial position $x \in \mathbb{R}$ and a target quantile $q \in (0, 1)$, the optimal deterministic schedule is given by an exponential schedule $X_t = X_0 e^{-t/\tau^\star}$ where the optimal time constant $\tau^\star$ is given by*

$$\tau^\star = \left(\frac{\eta x q}{\sqrt{2}\sigma\kappa(q)}\right)^{\frac{2}{3}}. \tag{4.44}$$

Let EXP *be this optimized exponential schedule. Then, its performance $V^{\mathrm{EXP}}(x, q)$ is given by*

$$V^{\mathrm{EXP}}(x, q) \triangleq \text{S-CVaR}_q\left[C_\infty^{x,\mathrm{EXP}}\right] = \frac{3}{2^{\frac{5}{3}}} \times \sigma^{\frac{2}{3}}\eta^{\frac{1}{3}} \times |x|^{\frac{4}{3}} \times q^{\frac{1}{3}}\big(\kappa(q)\big)^{\frac{2}{3}}. \tag{4.45}$$

While the proof is provided in Appendix B.1, we remark that the optimality of the exponential schedule can be directly inferred from the result of [49]: it was shown that, given a finite time-horizon of length T, a mean-variance optimization results in a trajectory $X_t = \frac{\sinh(\kappa(T-t))}{\sinh(\kappa T)} X_0$ for some constant $\kappa > 0$. As $T \nearrow \infty$, we can observe that the optimized trajectory converges to an exponential schedule $X_t = e^{-\kappa t} X_0$.

We next examine the volume-weighted average price (VWAP) schedules, under which the trader liquidates the asset at a constant rate until completion so that the trader's position decreases linearly over time.[10] The next proposition identifies the optimized VWAP schedule (see Appendix B.1 for the proof).

Proposition 4.5.2 (Optimized VWAP schedule)**.** *Given an initial position $x \in \mathbb{R}$ and a target quantile $q \in (0,1)$, the best VWAP schedule is given by $X_t = X_0\left(1 - \frac{t}{T^\star}\right)^+$, where the optimal execution period $T^\star$ is as follows:*

$$T^\star = \left(\frac{\sqrt{3}\eta x q}{\sigma\kappa(q)}\right)^{\frac{2}{3}}. \tag{4.46}$$

Let VWAP *be this optimized VWAP schedule. Then, its performance $V^{\mathrm{VWAP}}(x,q)$ is given by*

$$V^{\mathrm{VWAP}}(x,q) \triangleq \text{S-CVaR}_q\left[C_\infty^{x,\mathrm{VWAP}}\right] = \frac{3^{\frac{2}{3}}}{2} \times \sigma^{\frac{2}{3}}\eta^{\frac{1}{3}} \times |x|^{\frac{4}{3}} \times q^{\frac{1}{3}}\big(\kappa(q)\big)^{\frac{2}{3}}. \tag{4.47}$$

4.5.2 Cost analysis

We now compare three liquidation strategies: the optimal adaptive strategy (OPT) derived in §4.4, the optimized deterministic strategy (EXP), and the optimized VWAP strategy (VWAP). We have derived closed-form expressions (4.29), (4.45), and (4.47) that represent their S-CVaR performance V, V^{EXP}, and V^{VWAP}, respectively. Given that $V(x,q) \le V^{\mathrm{EXP}}(x,q) \le V^{\mathrm{VWAP}}(x,q)$ by their definitions, we particularly consider the following ratios that are useful for pairwise comparison:

$$\Upsilon_{\mathrm{OPT}}^{\mathrm{EXP}} \triangleq \frac{V^{\mathrm{EXP}}(x,q)}{V(x,q)} - 1, \quad \Upsilon_{\mathrm{OPT}}^{\mathrm{VWAP}} \triangleq \frac{V^{\mathrm{VWAP}}(x,q)}{V(x,q)} - 1, \quad \Upsilon_{\mathrm{EXP}}^{\mathrm{VWAP}} \triangleq \frac{V^{\mathrm{VWAP}}(x,q)}{V^{\mathrm{EXP}}(x,q)} - 1. \tag{4.48}$$

Note that these ratios do not change even if we compare CVaR performance instead of S-CVaR performance since S-CVaR is merely a scaled version of CVaR (see Remark 4.2.1.(i)).

These ratios can be expressed in closed form, and we observe the following. First, all the ratios

[10] A VWAP schedule typically refers to a strategy that liquidates an asset at a constant percent of volume (POV), e.g., 5% of market volume throughout the liquidation process. As we assume a time-stationary market, the market order flows are assumed stationary, and therefore a constant POV schedule is equivalent to a constant liquidation rate schedule.

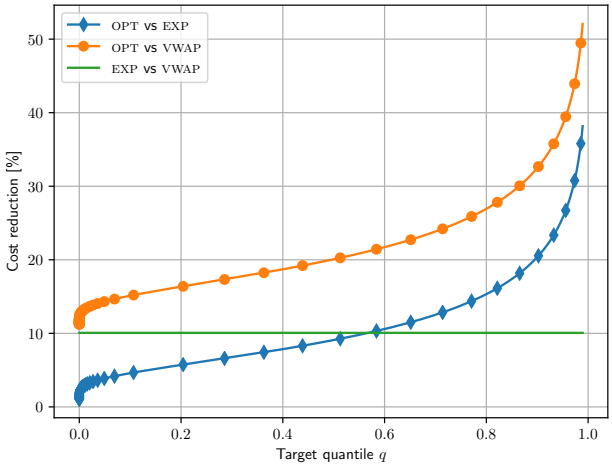

Figure 4.4: The pairwise comparison among three liquidation strategies – the optimal adaptive strategy, the optimal deterministic strategy (EXP), and the optimized VWAP strategy – measured with the ratios $\Upsilon^{\text{EXP}}_{\text{OPT}}$, $\Upsilon^{\text{VWAP}}_{\text{OPT}}$, and $\Upsilon^{\text{VWAP}}_{\text{EXP}}$ defined in (4.48). These ratios represent the relative percentage reductions in CVaR cost, and depend on the target quantile q only. Each curve shows the value of each ratio when q ranges from zero to one.

depend only on the quantile q, but not on the other problem parameters such as the quantity to liquidate x, the price volatility σ, and the market impact factor η. Figure 4.4 plots these ratios as functions of q. Second, the optimal deterministic schedule always outperforms to the best VWAP schedule by 10.0% (= $(4/3)^{1/3} - 1$), irrespective of the value of q. Finally, in a moderate range of q, from 0.2 to 0.8, the adaptive strategy outperforms the optimal deterministic strategy by 5% to 15%, and outperforms the optimized VWAP strategy by 15% to 25%. This gap increases as q increases (i.e., becomes more risk-neutral). More specifically, it diverges as q approaches one, and vanishes as q approaches zero.[11] See also §4.6.2 for a more illustrative comparison between the adaptive strategy and the deterministic strategies.

4.6 Numerical simulations

In this section, we provide the simulation results that confirm our theoretical findings and provide illustrative comparisons between the optimal strategy and the deterministic strategies. Hereafter, we denote the optimal adaptive strategy by OPT, the optimized deterministic schedule by

[11] The absolute performance (i.e., the CVaR cost) of all three policies converge to zero as $q \nearrow 1$, and diverge to infinity as $q \searrow 0$.

EXP, and the optimized VWAP schedule by VWAP.

4.6.1 Illustration of optimal adaptive strategy

We consider a situation where the trader wants to liquidate $x = 100$ units of an asset given the volatility $\sigma = 1.09$, the market impact factor $\eta = 0.0017$, and the target quantile q varying from 0.05 to 0.8. Without loss of generality, the initial price of the asset is set to zero.

We simulate the optimal policy OPT as follows. We first discretize the time horizon into subintervals of equal length $\Delta t = 10^{-4}$, and generate a sample path of the standard Brownian motion $W_0, W_{\Delta t}, W_{2\Delta t}, \ldots$. Starting from $X_0 = x$ and $Q_0 = q$, at each time $t = 0, \Delta t, 2\Delta t, \ldots$, we compute the liquidation rate $\pi_t = f^\star(X_t, Q_t)$ and the quantile diffusion rate $\gamma_t = g^\star(X_t, Q_t)$ and then update the position size $X_{t+\Delta t} = X_t - \pi_t \Delta t$ and the quantile $Q_{t+\Delta t} = Q_t + \gamma_t \Delta W_t$ accordingly, where $\Delta W_t \triangleq W_{t+\Delta t} - W_t$. The expressions for $f^\star$ and $g^\star$ are given in (4.36) and (4.37), and the value of $\varphi(q)$ can be computed using linear interpolation based on its parametric representation derived in Proposition 4.4.2. In order to prevent numerical instability, we keep the value of quantile process Q_t between ϵ and $1 - \epsilon$ via truncation (we take $\epsilon = 10^{-5}$); i.e., if $Q_t < \epsilon$ or $Q_t > 1 - \epsilon$, it is set to ϵ or $1 - \epsilon$, respectively. This procedure is repeated until the remaining position size X_t becomes smaller than 10^{-2}.

Figure 4.5 illustrates the sample paths of the price process σW_t, the position process X_t, and the quantile process Q_t under OPT for different values of target quantile $q \in \{0.05, 0.1, 0.2, 0.5, 0.8\}$ in the following two scenarios: when the price moves in an adverse direction (left), and when the price moves in a favorable direction (right). From these results, we confirm the behaviors of the optimal strategy characterized in §4.4.2. In every case, the position monotonically decreases over time; i.e., the optimal policy keeps trading in one direction. Also observe that the policy liquidates the position more aggressively as we take a smaller value for q (i.e., as the policy becomes more risk-averse). In a comparison between two scenarios (left vs. right), we observe "aggressiveness-in-the-money"; i.e., the policy trades more aggressively when the price moves in a favorable direction (right). This behavior can also be observed within each sample path: during

the execution process, the quantile process Q_t decreases when the price moves upward and the policy trades more aggressively. In addition, the quantile process Q_t converges to either zero or one, indicating whether the realized price process is among the worst q-fraction of the scenarios. While not reported here, we observe that Q_t converges to one in the q-fraction of simulations and converges to zero in the other $(1-q)$-fraction of simulations (recall that Q_t is a martingale starting at q).

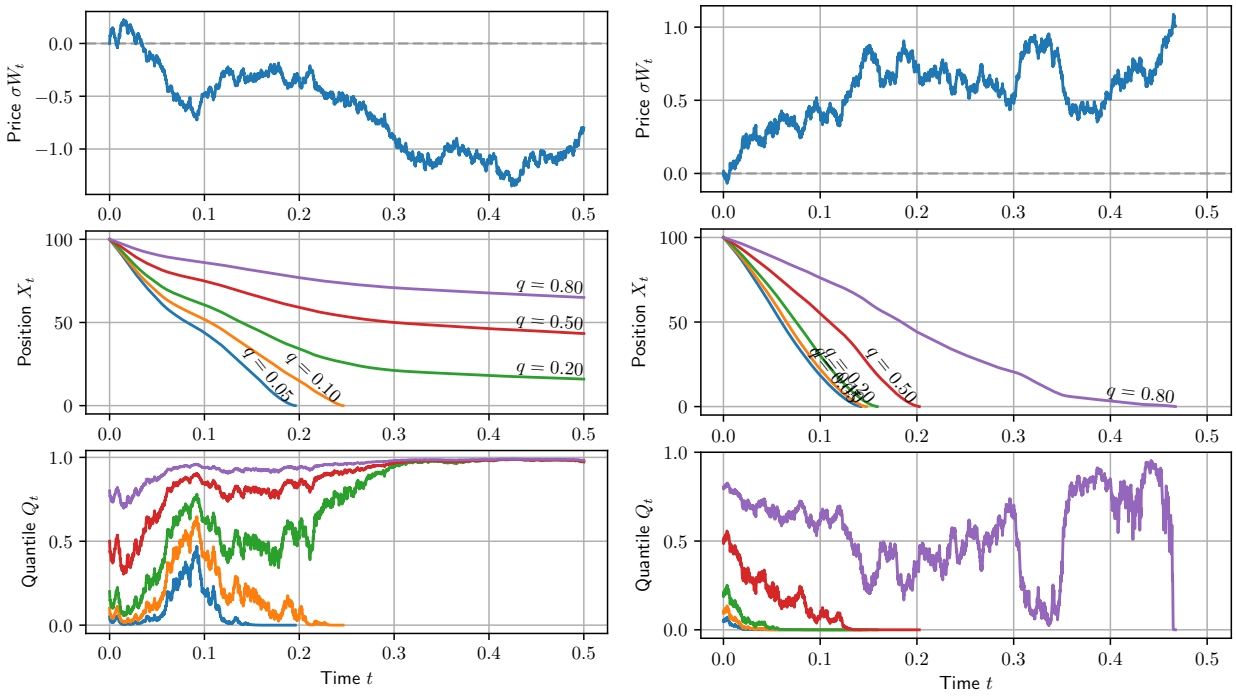

Figure 4.5: Illustration of the optimal adaptive liquidation processes with different values for target quantile $q \in \{0.05, 0.1, 0.2, 0.5, 0.8\}$ in two scenarios: when the price moves in an adverse direction (left), and when the price moves in a favorable direction (right). The plots in the top row show the realized price process over time, the plots in the middle row show the position processes X_t with a label on each curve indicating the target quantile of each strategy, and the plots in the bottom row show the quantile processes Q_t associated with these strategies.

4.6.2 Comparison with deterministic strategies

We provide the simulation results of the deterministic strategies (EXP, VWAP) introduced in §4.5.1 in a comparison with those of the optimal adaptive strategy (OPT). Figure 4.6 shows the position process trajectories under these three strategies with target quantile $q = 0.5$ in the two

scenarios as in Figure 4.5. One can immediately observe that the deterministic strategies are not adaptive to the price changes. The optimal adaptive strategy liquidates at a similar rate to the exponential schedule during the initial periods, but it deviates as soon as it adjusts its aggressiveness adaptively to the price changes. In particular, it becomes more aggressive when the price moves in its favor.

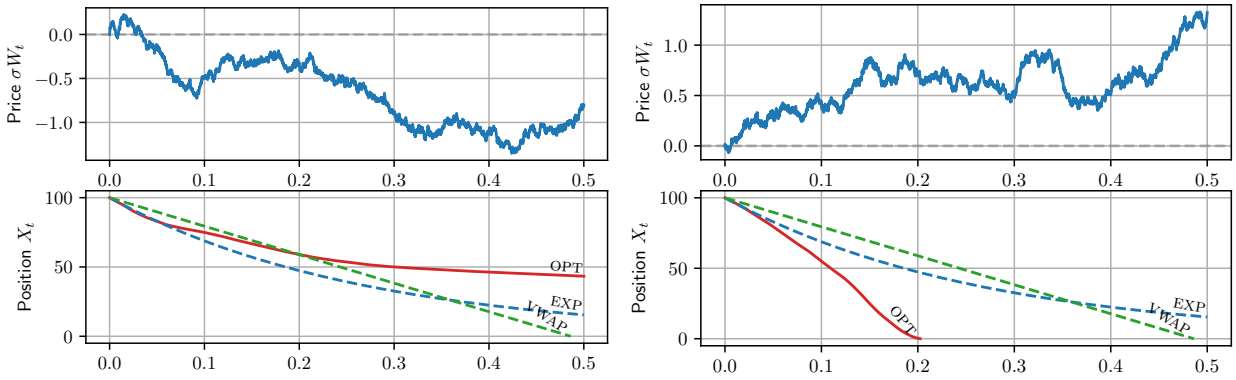

Figure 4.6: Illustration of liquidation processes under the optimal adaptive strategy (OPT, red solid lines), the optimized deterministic schedule (EXP, blue dashed lines), and the optimized VWAP schedule (VWAP, green dashed lines) with the target quantile $q = 0.5$, and in two scenarios: when the price moves in an adverse direction (left), and when the price moves in a favorable direction (right).

Figure 4.7 shows the implementation shortfall distributions (i.e., the histograms of C_∞) resulting from OPT (top) and EXP (bottom) with different values of the target quantile $q \in \{0, 1, 0.2, 0.5\}$. These histograms are obtained from 100,000 runs of simulations, where all the strategies see the same price process realization per simulation. The resulting distributions are visually very different: EXP yields a normal distribution whereas OPT yields a distribution that has a sharp peak at the q^{th} quantile. Such a sharp peak can be explained by the threshold behavior of the optimal adaptive strategy, discussed at the end of §4.4.2.

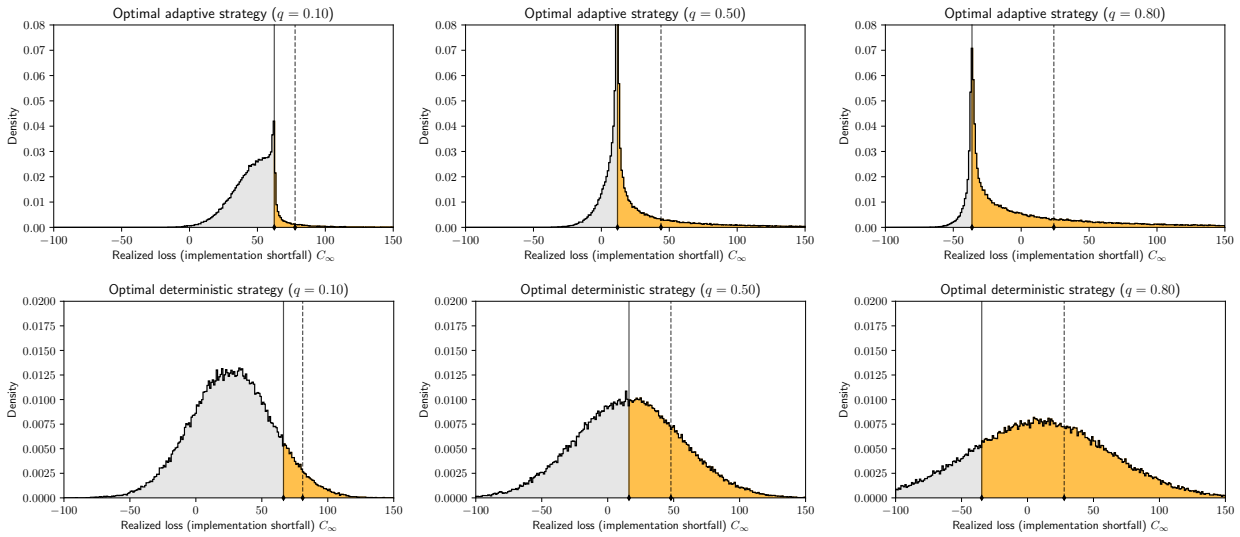

Figure 4.7: The distributions of the implementation shortfall (i.e., the histogram of C_∞) incurred by the optimal adaptive liquidation strategy (top) and the optimized deterministic schedule (bottom), given the initial position $x = 100$, the volatility $\sigma = 1.09$, the market impact factor $\eta = 0.017$, and a varying target quantile $q \in \{0.1, 0.5, 0.8\}$ (left, middle, and right, respectively). In each plot, the solid vertical line represents the q^{th} quantile (i.e., the value-at-risk $\text{VaR}_q[C_\infty]$), and the dashed vertical line represents the tail average beyond the q^{th} quantile (i.e., the conditional risk-at-value $\text{CVaR}_q[C_\infty]$). The highlighted area represents the worst q-fraction of the outcomes whose average corresponds to $\text{CVaR}_q[C_\infty]$. These are obtained from 100,000 runs of simulations.

Table 4.1 and Figure 4.8 report the summary statistics of the implementation shortfall obtained from those simulations. We confirm that the simulation results are consistent with our theoretic predictions, and the derived optimal adaptive strategy effectively reduces the CVaR loss across all settings.

Target quantile	Policy	CVaR (theo.)	VaR	Average	Std. dev.
	OPT	87.10 (86.41)	74.20	52.80	17.63
$q = 0.05$	EXP	90.15 (89.77)	77.98	30.11	29.01
	VWAP	99.18 (98.80)	85.99	33.13	31.92
	OPT	77.74 (77.07)	62.41	47.56	18.66
$q = 0.10$	EXP	80.97 (80.60)	66.43	27.06	30.61
	VWAP	89.14 (88.71)	73.09	29.77	33.68
	OPT	66.03 (65.58)	46.95	40.08	22.77
$q = 0.20$	EXP	69.62 (69.32)	50.98	23.31	33.01
	VWAP	76.64 (76.30)	56.14	25.64	36.32
	OPT	43.90 (43.68)	11.88	24.29	42.74
$q = 0.50$	EXP	47.92 (47.66)	16.14	16.12	39.84
	VWAP	52.70 (52.45)	17.65	17.70	43.82
	OPT	24.05 (23.85)	−36.23	11.22	81.33
$q = 0.80$	EXP	27.81 (27.51)	−34.70	9.45	52.49
	VWAP	30.65 (30.28)	−38.21	10.42	57.80

Table 4.1: Summary statistics of implementation shortfall C_∞ incurred by the optimal adaptive strategy (OPT), the optimized deterministic schedule (EXP), and the optimized VWAP schedule (VWAP), given $x = 100$, $\sigma = 1.09$, $\eta = 0.017$, and a varying target quantile $q \in \{0.05, 0.1, 0.2, 0.5, 0.8\}$. For each combination of a policy and a target quantile, it reports the CVaR value, the VaR value, the average, and the standard deviation of implementation shortfall C_∞, measured from 100,000 runs of simulations. The numbers in parentheses in the third column report the theoretically predicted CVaR values, computed with the expressions (4.29), (4.45), and (4.47).

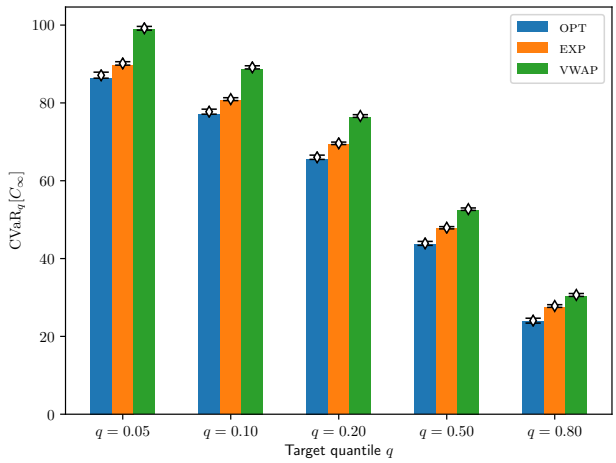

Figure 4.8: Comparison of CVaR performance achieved by the optimal adaptive strategy (OPT), the optimized deterministic schedule (EXP), and the optimized VWAP schedule (VWAP), given $x = 100$, $\sigma = 1.09$, $\eta = 0.017$, and a varying target quantile $q \in \{0.05, 0.1, 0.2, 0.5, 0.8\}$. Each data point (plotted with a diamond-shaped marker) reports the CVaR value of the implementation shortfall distribution obtained from 100,000 runs of simulations, where the error bar around the data point represents a 90% confidence interval computed via bootstrapping. The colored bars represent their theoretical predictions computed with the expressions (4.29), (4.45), and (4.47).

References

[1] J. C. Gittins, "Bandit processes and dynamic allocation indices," *Journal of the Royal Statistical Society, Series B*, vol. 41, no. 2, pp. 148–177, 1979.

[2] R. N. Bradt, S. M. Johnson, and S. Karlin, "On sequential designs for maximizing the sum of *n* observations," *Annals of Mathematical Statistics*, vol. 27, no. 4, pp. 1060–1074, 1956.

[3] D. A. Berry and B. Fristedt, *Bandit Problems: Sequential Allocation of Experiments*. Chapman and Hall, 1985.

[4] D. B. Brown, J. E. Smith, and P. Sun, "Information relaxations and duality in stochastic dynamic programs," *Operations Research*, vol. 58, no. 4, pp. 785–801, 2010.

[5] R. T. Rockafellar and R. J.-B. Wets, "Scenarios and policy aggregation in optimization under uncertainty," *Mathematics of Operations Research*, vol. 16, no. 1, pp. 119–147, 1991.

[6] M. H. A. Davis and I. Karatzas, *A Deterministic Approach to Optimal Stopping*. Wiley, 1994.

[7] L. C. G. Rogers, "Monte Carlo valuation of American options," *Mathematical Finance*, vol. 12, no. 3, pp. 271–286, 2002.

[8] M. B. Haugh and L. Kogan, "Pricing American options: A duality approach," *Operations Research*, vol. 52, no. 2, pp. 258–270, 2004.

[9] V. V. Desai, V. F. Farias, and C. C. Moallemi, "Pathwise optimization for optimal stopping problems," *Management Science*, vol. 58, no. 12, pp. 2292–2308, 2012.

[10] ——, "Bounds for Markov decision processes," in *Reinforcement Learning and Approximate Dynamic Programming for Feedback Control*, F. L. Lewis and D. Liu, Eds., 2012, pp. 452–473.

[11] M. B. Haugh and A. E. B. Lim, "Linear-quadratic control and information relaxations," *Operations Research Letters*, vol. 40, pp. 521–528, 2012.

[12] M. B. Haugh and C. Wang, "Dynamic portfolio execution and information relaxations," *SIAM Journal of Financial Math*, vol. 5, pp. 316–359, 2014.

[13] D. B. Brown and M. B. Haugh, "Information relaxation bounds for infinite horizon Markov decision processes," *Operations Research*, vol. 65, no. 5, pp. 1355–1379, 2017.

[14] M. B. Haugh and O. R. Lacedelli, "Information relaxation bounds for partially observed Markov decision processes," *IEEE Transactions on Automatic Control*, 2019.

[15] W. Thompson, "On the likelihood that one unknown probability exceeds another in view of the evidence of two samples," *Biometrika*, vol. 25, no. 3/4, pp. 285–294, 1933.

[16] T. L. Lai and H. Robbins, "Asymptotically efficient adaptive allocation rules," *Advances in Applied Mathematics*, vol. 6, pp. 4–22, 1985.

[17] E. Kaufmann, N. Korda, and R. Munos, "Thompson sampling: An asymptotically optimal finite-time analysis," in *Algorithmic Learning Theory*, ser. Lecture Notes in Computer Science, N. Bshouty, G. S. G., N. Vayatis, and T. Zeugmann, Eds., vol. 7568, Springer, 2012.

[18] S. Agrawal and N. Goyal, "Further optimal regret bounds for Thompson sampling," *Proceeds of the 16th International Conference on Artificial Intelligence and Statistics*, pp. 99–107, 2013.

[19] S. Bubeck and C.-Y. Liu, "Prior-free and prior-dependent regret bounds for Thompson sampling," *Proceedings of the 26th International Conference on Neural Information Processing Systems*, vol. 1, no. 638–646, 2013.

[20] E. Gutiérrez-Peña and A. F. M. Smith, "Conjugate parameterizations for natural exponential families," *Journal of the American Statistical Association*, vol. 90, no. 432, pp. 1347–1356, 2012.

[21] D. Russo and B. Van Roy, "Learning to optimize via posterior sampling," *Mathematics of Operations Research*, vol. 39, no. 4, pp. 1221–1243, 2014.

[22] ——, "Learning to optimize via information-directed sampling," *Operations Research*, vol. 66, no. 1, pp. 230–252, 2017.

[23] E. Kaufmann, O. Cappé, and A. Garivier, "On Bayesian upper confidence bounds for bandit problems," *Proceedings of the Fifteenth International Conference on Artificial Intelligence and Statistics,*, vol. 22, pp. 592–600, 2012.

[24] J. Niño-Mora, "Computing a classic index for finite-horzion bandits," *INFORMS Journal on Computing*, vol. 23, no. 2, pp. 254–267, 2011.

[25] T. Lattimore, "Regret analysis of the finite-horizon Gittins index strategy for multi-armed bandits," *29th Annual Conference on Learning Theory*, vol. 49, pp. 1–32, 2016.

[26] V. F. Farias and E. Gutin, "Optimistic Gittins indices," *Proceedings of the 30th International Conference on Neural Information Processing Systems*, pp. 3161–3169, 2016.

[27] D. B. Brown and J. E. Smith, "Index policies and performance bounds for dynamic selection problems," *Management Science*, vol. 66, no. 7, pp. 3029–3050, 2020.

[28] W. Ding, T. Qin, X.-D. Zhang, and T.-Y. Liu, "Multi-armed bandit with budget constraint and variable costs," *Proceedings of the 27th AAAI Conference on Artificial Intelligence*, 2013.

[29] D. Russo, D. Tse, and B. Van Roy, "Time-sensitive bandit learning and satisficing Thompson sampling," 2017.

[30] D. Russo, B. Van Roy, A. Kazerouni, I. Osband, and Z. Wen, "A tutorial on Thompson sampling," *Foundations and Trends in Machine Learning*, vol. 11, no. 1, pp. 1–96, 2018.

[31] R. J. Williams, "Simple statistical gradient-following algorithms for connectionist reinforcement learning," *Machine Learning*, vol. 8, no. 3-4, pp. 229–256, 1992.

[32] C. Finn, P. Abbeel, and S. Levine, "Model-agnostic meta-learning for fast adaptation of deep networks," in *International Conference on Machine Learning*, 2017, pp. 1126–1135.

[33] J. Baxter, "A model of inductive bias learning," *Journal of Artificial Intelligence Research*, vol. 12, pp. 149–198, 2000.

[34] H. Bastani, D. Simchi-Levi, and R. Zhu, "Meta dynamic pricing: Learning across experiments," *Available at SSRN 3334629*, 2019.

[35] Y. Duan, J. Schulman, X. Chen, P. L. Bartlett, I. Sutskever, and P. Abbeel, "RL^2: Fast reinforcement learning via slow reinforcement learning," *arXiv preprint arXiv:1611.02779*, 2016.

[36] C. Boutilier, C.-W. Hsu, B. Kveton, M. Mladenov, C. Szepesvari, and M. Zaheer, "Differentiable bandit exploration," *arXiv preprint arXiv:2002.06772*, 2020.

[37] B. Kveton, M. Mladenov, C.-W. Hsu, M. Zaheer, C. Szepesvari, and C. Boutilier, "Differentiable meta-learning in contextual bandits," *arXiv preprint arXiv:2006.05094*, 2020.

[38] E. Gutin and V. Farias, "Optimistic Gittins indices," in *Advances in Neural Information Processing Systems*, 2016, pp. 3153–3161.

[39] D. Russo and B. Van Roy, "Learning to optimize via information-directed sampling," *Operations Research*, vol. 66, no. 1, pp. 230–252, 2018.

[40] E. Kaufmann, O. Cappé, and A. Garivier, "On Bayesian upper confidence bounds for bandit problems," in *Artificial Intelligence and Statistics*, 2012, pp. 592–600.

[41] W. R. Thompson, "On the likelihood that one unknown probability exceeds another in view of the evidence of two samples," *Biometrika*, vol. 25, no. 3/4, pp. 285–294, 1933.

[42] O. Chapelle and L. Li, "An empirical evaluation of Thompson sampling," in *Advances in Neural Information Processing Systems*, 2011, pp. 2249–2257.

[43] S. Min, C. Maglaras, and C. C. Moallemi, "Thompson sampling with information relaxation penalties," Working paper, 2020.

[44] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," *arXiv preprint arXiv:1412.6980*, 2014.

[45] P. L'Ecuyer, "Note: On the interchange of derivative and expectation for likelihood ratio derivative estimators," *Management Science*, vol. 41, no. 4, pp. 738–747, 1995.

[46] R. S. Sutton and A. G. Barto, *Reinforcement Learning: An Introduction*. MIT Press, 2018.

[47] P. Glasserman and D. D. Yao, "Some guidelines and guarantees for common random numbers," *Management Science*, vol. 38, no. 6, pp. 884–908, 1992.

[48] D. Russo and B. Van Roy, "Satisficing in time-sensitive bandit learning," *arXiv preprint arXiv:1803.02855*, 2018.

[49] R. Almgren and N. Chriss, "Optimal execution of portfolio transactions," 2000.

[50] R. Kissell and R. Malamut, "Understanding the profit and loss distribution of trading algorithms," in *Algorithmic Trading*, B. R. Bruce, Ed., Institutional Investor, 2005, pp. 41–49.

[51] J. Lorenz and R. Almgren, "Adaptive arrival price," in *Algorithmic Trading III*, B. R. Bruce, Ed., Institutional Investor, 2007, pp. 59–66.

[52] ——, "Mean-variance optimal adaptive execution," *Applied Mathematical Finance*, vol. 18, no. 5, pp. 395–422, 2011.

[53] P. A. Forsyth, "A Hamilton–Jacobi–Bellman approach to optimal trade execution," *Applied Numerical Mathematics*, vol. 61, pp. 241–265, 2011.

[54] A. Schied and T. Schöneborn, "Risk aversion and the dynamics of optimal liquidation strategies in illiquid markets," *Finance and Stochastics*, vol. 13, no. 2, pp. 181–204, 2009.

[55] J. Gatheral and A. Schied, "Optimal trade execution under geometric Brownian motion in the Almgren and Chriss framework," *International Journal of Theoretical and Applied Finance*, vol. 14, no. 3, pp. 353–368, 2011.

[56] P. A. Forsyth, S. Kennedy, S. T. Tse, and H. Windcliff, "Optimal trade execution: A mean quadratic variation approach," *Journal of Economic Dynamics & Control*, vol. 36, pp. 1971–1991, 2012.

[57] Q. Lin, X. Chen, and J. Peña, "A trade execution model under a composite dynamic coherent risk measure," *Operations Research Letters*, vol. 43, pp. 52–58, 2015.

[58] P. Glasserman and X. Xu, "Robust portfolio control with stochastic factor dynamics," *Operations Research*, vol. 61, no. 4, pp. 874–893, 2013.

[59] P. Artzner, F. Delbaen, J.-M. Eber, and D. Heath, "Coherent measures of risk," *Mathematical Finance*, vol. 9, no. 3, pp. 203–228, 1999.

[60] R. T. Rockafellar and S. Uryasev, "Conditional value-at-risk for general loss distributions," *Journal of Banking & Finance*, vol. 26, pp. 1443–1471, 2002.

[61] P. Artzner, F. Delbaen, J.-M. Eber, D. Heath, and H. Ku, "Coherent multiperiod risk adjusted values and Bellman's principle," *Annals of Operations Research*, vol. 152, pp. 5–22, 2007.

[62] A. Shapiro, "On a time consistency concept in risk averse multistage stochastic programming," *Operations Research Letters*, vol. 37, pp. 143–147, 2009.

[63] N. Bäuerle and J. Ott, "Markov decision processes with average-value-at-risk criteria," *Mathematical Methods of Operational Research*, vol. 74, no. 4, pp. 361–379, 2011.

[64] Y. Huang and X. Guo, "Minimum average value-at-risk for finite horizon semi-Markov decision processes in continuous time," *SIAM Journal on Optimization*, vol. 26, no. 1, pp. 1–28, 2016.

[65] C. W. Miller and I. Yang, "Optimal control of conditional value-at-risk in continuous time," *SIAM Journal on Control and Optimization*, vol. 55, no. 2, pp. 856–884, 2017.

[66] Y. Chow, M. Ghavamzadeh, L. Janson, and M. Pavone, "Risk-constrained reinforcement learning with percentile risk criteria," *Journal of Machine Learning Research*, vol. 18, pp. 1–51, 2018.

[67] J. Backhoff-Veraguas and L. Tangpi, "On the dynamic representation of some time-inconsistent risk measures in a Brownian filtration," *Mathematics and Financial Economics*, vol. 14, pp. 433–460, 2020.

[68] G. C. Pflug and A. Pichler, "Time-inconsistent multistage stochastic programs: Martingale bounds," *European Journal of Operational Research*, vol. 249, pp. 155–163, 2016.

[69] Y. Chow, A. Tamar, S. Mannor, and M. Pavone, "Risk-sensitive and robust decision-making: A CVaR optimization approach," *Advances in Neural Information Processing Systems 28 (NIPS 2015)*, 2015.

[70] M. P. Chapman, J. P. Lacotte, K. M. Smith, I. Yang, Y. Han, M. Pavone, and C. J. Tomlin, "Risk-sensitive safety specifications for stochastic system using conditional value-at-risk," 2018.

[71] X. Li, H. Zhong, and M. L. Brandeau, "Quantile Markov decision processes," 2020.

[72] P. Protter, *Stochastic Integration and Differential Equations*. Springer, 2015.

[73] M. Sion, "On general minimax theorems," *Pacific Journal of Mathematics*, vol. 8, no. 1, pp. 171–176, 1958.

[74] A. D. Polyanin and V. F. Zaitsev, *Handbook of Exact Solutions for Ordinary Differential Equations*. Chapman & Hall/CRC Press, 2003.

[75] R. Durrett, *Probability: Theory and Examples*, 4th. Cambridge University Press, 2010.

[76] V. Barbu and T. Precupanu, *Convexity and Optimization in Banach Spaces*, 4th. Springer, 2012.

Appendix A: Appendix for Thompson Sampling with Information Relaxation Penalties

A.1 An illustrative example

Let us consider a Bernoulli MAB with eight periods ($T = 8$) and three arms ($K = 3$) with the following priors:

$$\mu_1 \sim \text{Beta}(3,1), \quad \mu_2 \sim \text{Beta}(1,1), \quad \mu_3 \sim \text{Beta}(1,3), \tag{A.1}$$

where $R_{a,n} \sim \text{Bernoulli}(\mu_a)$ for each $a \in \{1,2,3\}$ and $n \in \{1,2,\cdots,8\}$. Given this prior belief, the predictive mean reward of each arm is $\bar{\mu}_1 = \mathbb{E}_{\mu_1 \sim \text{Beta}(3,1)}[\mu_1] = \frac{3}{4}$, $\bar{\mu}_2 = \frac{1}{2}$, and $\bar{\mu}_3 = \frac{1}{4}$, respectively. As an illustrative example, we examine a particular instance where the true outcome ω is given as follows:

	True means $\mu_a(\theta_a)$	Rewards $R_{a,n}$							
		$n=1$	2	3	4	5	6	7	8
Arm 1 ($a=1$)	0.235	0	1	1	1	0	0	0	0
Arm 2 ($a=2$)	0.443	1	0	0	1	1	1	1	0
Arm 3 ($a=3$)	0.787	1	1	1	1	0	0	1	1

Table A.1: An example of the outcome in a Bernoulli MAB with $K = 3$ and $T = 8$.

If we consider only the priors, arm 1 is best since $\bar{\mu}_1$ is largest among $(\bar{\mu}_1, \bar{\mu}_2, \bar{\mu}_3)$. If, however, we have full information about the parameter values, arm 3 is best since μ_3 is largest among (μ_1, μ_2, μ_3).

A.1.1 Inner Problems Induced by Different Penalty Functions

No penalty. To clarify the role of penalties, we first consider the case of zero penalty, i.e., $z_t \equiv 0$, which was not discussed in §2.3. With zero penalty, the DM at any time earns the current realized reward without adjustment. The clairvoyant DM, who is informed of the outcome ω, can find the best action sequence for this particular outcome ω. Recall that $R_{a,n}$ is defined to be the reward from the n^{th} pull of arm a, not the reward from arm a at time n, and so the DM is not allowed to skip any of the reward realizations and the total reward does not depend on the order of pulls. As depicted in the table below, the optimal solution is to pull arm 1 four times, arm 2 once, and arm 3 three times, which yields a total reward of 7.

	Payoffs under zero penalty								Maximal payoff
	$n = 1$	2	3	4	5	6	7	8	
Arm 1	0	1	1	1	0	0	0	0	
Arm 2	1	0	0	1	1	1	1	0	7
Arm 3	1	1	1	1	0	0	1	1	

TS penalty. Next, let us examine the penalty $z_t^{\text{TS}}(\mathbf{a}_{1:t}, \omega) \triangleq r_t(\mathbf{a}_{1:t}, \omega) - \mu_{a_t}(\theta_{a_t})$ under which the DM earns μ_a whenever playing arm a. The hindsight optimal action sequence is to pull arm 3 (the arm with the largest mean reward μ_a) eight times in a row and the DM can earn a total reward of $T \times \mu_3 = 6.296$ at most.

	Payoffs under z_t^{TS}								Maximal payoff
	$n = 1$	2	3	4	5	6	7	8	
Arm 1	.235	.235	.235	.235	.235	.235	.235	.235	
Arm 2	.443	.443	.443	.443	.443	.443	.443	.443	6.296
Arm 3	.787	.787	.787	.787	.787	.787	.787	.787	

IRS.FH penalty. When the penalties are given by $z_t^{\text{IRS.FH}}(\mathbf{a}_{1:t}, \omega) \triangleq r_t(\mathbf{a}_{1:t}, \omega) - \hat{\mu}_{a_t,T-1}(\omega)$, the DM earns $\hat{\mu}_{a,T-1}(\omega)$ whenever playing arm a. Recall that $\hat{\mu}_{a,T-1}(\omega)$ is the Bayesian estimate on mean reward of arm a after observing reward realizations $R_{a,1}, \cdots, R_{a,T-1}$. In this particular example, we have $\left(\hat{\mu}_{1,T-1}, \hat{\mu}_{2,T-1}, \hat{\mu}_{3,T-1}\right) = \left(\frac{6}{11}, \frac{6}{9}, \frac{6}{11}\right)$ and the maximal payoff is $T \times \hat{\mu}_{2,T-1} = 5.333$, which can be obtained by playing arm 2 throughout the entire time horizon.

	Payoffs under $z_t^{\text{IRS.FH}}$								Maximal payoff
	$n=1$	2	3	4	5	6	7	8	
Arm 1	6/11	6/11	6/11	6/11	6/11	6/11	6/11	6/11	
Arm 2	6/9	6/9	6/9	6/9	6/9	6/9	6/9	6/9	5.333
Arm 3	6/11	6/11	6/11	6/11	6/11	6/11	6/11	6/11	

IRS.V-ZERO penalty. Finally, let us focus on $z_t^{\text{IRS.V-ZERO}}(\mathbf{a}_{1:t}, \omega) \triangleq r_t(\mathbf{a}_{1:t}, \omega) - \hat{\mu}_{a_t, n_{t-1}(\mathbf{a}_{1:t-1}, a_t)}$ under which the DM earns $\hat{\mu}_{a,n-1}(\omega)$ from the n^{th} pull of arm a. Since the payoff from an arm changes over time as the Bayesian estimate evolves, playing only one arm is no longer optimal, unlike in the previous two cases. It can be easily verified that the optimal allocation is to play arm 1 six times and arm 2 two times, as visualized in the table below.

	Payoffs under $z_t^{\text{IRS.V-ZERO}}$								Maximal payoff
	$n=1$	2	3	4	5	6	7	8	
Arm 1	3/4	3/5	4/6	5/7	6/8	6/9	6/10	6/11	
Arm 2	1/2	2/3	2/4	2/5	3/6	4/7	5/8	6/9	5.314
Arm 3	1/4	2/5	3/6	4/7	5/8	5/9	5/10	6/11	

IRS.V-EMAX and the ideal penalty. Regarding the penalty functions $z_t^{\text{IRS.V-EMAX}}$ and z_t^{ideal}, we cannot visualize the optimal solution with a table since the total payoff depends on the detailed sequence of pulls and not only the number of pulls. While omitting the visual proof of optimality, we have that the action sequence $\mathbf{a}^*_{1:8} = (1, 2, 2, 1, 1, 1, 1, 1)$ achieves the maximal payoff of 5.806 under $z_t^{\text{IRS.V-EMAX}}$, and $\mathbf{a}^*_{1:8} = (1, 1, 1, 1, 1, 1, 1, 1)$ achieves the maximal payoff of 6.063 under z_t^{ideal}. In particular for z_t^{ideal}, the maximal payoff depends only on the prior belief $\mathbf{y}$ and the time horizon T, irrespective of the outcome[1] ω.

We have so far illustrated how the different penalty functions induce the different inner problems and the different best actions given the same outcome ω. The readers may notice from the above examples that, as the penalty function becomes more complicated, the hindsight best action sequence becomes less dependent on a particular realization of ω. Instead, it becomes more

[1]For details, see the proof of the strong duality theorem in §A.3.1. While the maximal value does not depend on ω, the optimal action sequence still depends on ω. More specifically, it is the sequence of actions that the (non-anticipating) Bayesian optimal policy will take when ω is sequentially revealed.

dependent on the prior belief.

A.1.2 IRS Performance Bounds

The maximal payoffs above are calculated for a particular outcome given by Table A.1. Recall that the IRS performance bound W^z is defined as the expected value of the maximal payoff where the expectation is taken with respect to the randomness of outcome ω over its prior distribution $\mathcal{I}(T, \mathbf{y})$. We can obtain this value by simulation, i.e., by solving a bunch of inner problems with respect to the randomly generated outcomes $\omega^{(1)}, \omega^{(2)}, \cdots, \omega^{(S)}$ and taking the average of the maximal values. For this particular Bernoulli MAB setting ($T = 8$ with given priors), we obtain the following performance bounds:

W^0	W^{TS}	$W^{\text{IRS.FH}}$	$W^{\text{IRS.V-ZERO}}$	$W^{\text{IRS.V-EMAX}}$	$W^{\text{ideal}} = V^*$
6.805	6.429	6.279	6.111	6.075	6.063

We observe that the performance bounds are monotone, i.e., $W^0 > W^{\text{TS}} > W^{\text{IRS.FH}} > W^{\text{IRS.V-ZERO}} > W^{\text{IRS.V-EMAX}} > W^{\text{ideal}} = V^*$, which is consistent with Theorem 2.4.1.

A.1.3 Illustration of the IRS Policy (IRS.V-Zero)

We illustrate how the policy $\pi^{\text{IRS.V-ZERO}}$ makes decisions sequentially when the true outcome ω is the one specified in Table A.1. At $t = 1$, it first synthesizes a future scenario based on the prior belief (i.e., sampling $\tilde{\omega}_1 \sim \mathcal{I}(\mathbf{y}_0)$) and finds the best action sequence in the presence of penalties $z_t^{\text{IRS.V-ZERO}}$ in the belief that the sampled outcome $\tilde{\omega}_1$ is the ground truth. The following table shows an example in which $\pi^{\text{IRS.V-ZERO}}$ plays arm 1.

$t = 1$	Priors $\mathbf{y}_0$	Payoffs with respect to $\tilde{\omega}_1 \sim \mathcal{I}(\mathbf{y}_0)$								Action
		$n = 1$	2	3	4	5	6	7	8	
Arm 1	Beta(3, 1)	3/4	4/5	5/6	6/7	7/8	7/9	8/10	9/11	
Arm 2	Beta(1, 1)	1/2	1/3	1/4	1/5	1/6	1/7	2/8	3/9	$a_1 = 1$
Arm 3	Beta(1, 3)	1/4	1/5	1/6	1/7	1/8	1/9	1/10	2/11	

As a result of the first action ($a_1 = 1$), we observe that $R_{1,1} = 0$ (encoded in the true outcome ω)

and the associated belief is updated from Beta(3, 1) to Beta(3, 2) according to Bayes' rule. In order to make the next decision a_2 at time $t = 2$, $\pi^{\text{IRS.V-ZERO}}$ simulates an outcome for the remaining time horizon, i.e., $\tilde{\omega}_2 \sim \mathcal{I}(\mathbf{y}_1)$, independently of the outcome $\tilde{\omega}_1$ used at $t = 1$. Again, $\pi^{\text{IRS.V-ZERO}}$ finds the best action sequence for this new scenario and takes its first action.[2] The table below shows an instance of $\tilde{\omega}_2$ in which the policy will pull arm 2.

$t = 2$	Priors $\mathbf{y}_1$	Payoffs with respect to $\tilde{\omega}_2 \sim \mathcal{I}(\mathbf{y}_1)$							Action
		$n = 1$	2	3	4	5	6	7	
Arm 1	Beta(3, 2)	3/5	4/6	4/7	4/8	4/9	5/10	5/11	
Arm 2	Beta(1, 1)	1/2	2/3	3/4	3/5	4/6	4/7	5/8	$a_2 = 2$
Arm 3	Beta(1, 3)	1/4	1/5	1/6	1/7	1/8	1/9	1/10	

We can update the prior of arm 2 as a new reward realization $R_{2,1} = 1$ is revealed. In the following decision epochs $t = 3, 4, \cdots$, the policy repeats the same decision-making procedure – (i) samples $\tilde{\omega}_t \sim \mathcal{I}(\mathbf{y}_{t-1})$, (ii) solves the inner problem, and (iii) plays the best arm that the optimal solution suggests – while updating the priors as the true reward realizations are revealed sequentially.

The following table illustrates the last decision epoch. As there remains one time period only, the policy $\pi^{\text{IRS.V-ZERO}}$ tries to maximize $\hat{\mu}_{a,0}(\tilde{\omega}_7) = \bar{\mu}_a(\mathbf{y}_7)$, which is the expected mean reward given the prior at that moment. Such a decision is totally myopic, but it is Bayesian optimal.

$t = 8$	Priors $\mathbf{y}_7$	Payoffs with respect to $\tilde{\omega}_7 \sim \mathcal{I}(\mathbf{y}_7)$	Action
		$n = 1$	
Arm 1	Beta(6, 3)	6/9	
Arm 2	Beta(2, 2)	2/4	$a_8 = 1$
Arm 3	Beta(1, 3)	1/4	

[2] In case of IRS.V-ZERO, we select the arm with the largest pull allocation as a first action.

A.2 Algorithms in detail

A.2.1 Implementation of IRS.V-ZERO

We provide a pseudo-code of the policy $\pi^{\text{IRS.V-ZERO}}$ introduced in §2.3.3. The same logic can be directly used to compute the performance bound $W^{\text{IRS.V-ZERO}}$ if the sampled outcome $\tilde{\omega}$ is replaced

with the true outcome ω.

Algorithm 5: Arm selection rule of $\pi^{\text{IRS.V-ZERO}}$ when remaining time is T and current belief is $\mathbf{y}$

Function `IRS.V-Zero(`$T, \mathbf{y}$`)`

$\tilde{\theta}_a \sim \mathcal{P}_a(y_a), \tilde{R}_{a,n} \sim \mathcal{R}_a(\tilde{\theta}), \quad \forall n \in \{1, \ldots, T\}, \forall a \in \{1, \ldots, K\}$

for $a = 1, \cdots, K$ **do**

  $\tilde{y}_{a,0} \leftarrow y_a, \tilde{S}_{a,0} \leftarrow 0$

  for $n = 1, \cdots, T$ **do**

    $\tilde{S}_{a,n} \leftarrow \tilde{S}_{a,n-1} + \bar{\mu}_a(\tilde{y}_{a,n-1})$

    $\tilde{y}_{a,n} \leftarrow \mathcal{U}_a(\tilde{y}_{a,n-1}, \tilde{R}_{a,n})$

  end

end

$\tilde{M}_{0,0} \leftarrow 0, \tilde{M}_{0,n} \leftarrow -\infty, \forall n \in \{1, \ldots, T\}$

for $a = 1, \cdots, K$ **do**

  for $n = 0, \cdots, T$ **do**

    $\tilde{M}_{a,n} \leftarrow \max_{0 \le m \le n}\{\tilde{M}_{a-1,n-m} + \tilde{S}_{a,m}\}$

    $\tilde{L}_{a,n} \leftarrow \operatorname{argmax}_{0 \le m \le n}\{\tilde{M}_{a-1,n-m} + \tilde{S}_{a,m}\}$

  end

end

$\tau \leftarrow T$

for $a = K, \cdots, 1$ **do**

  $\tilde{n}_a^* \leftarrow \tilde{L}_{a,\tau}$

  $\tau \leftarrow \tau - \tilde{n}_a^*$

end

return $\operatorname{argmax}_a \tilde{n}_a^*$

A.2.2 Implementation of IRS.V-EMAX

We use the notation $\mathbf{y}_t(\mathbf{n}_{1:K}, \omega)$ to denote the belief as a function of pull counts $\mathbf{n}_{1:K} \triangleq (n_1, \cdots, n_K) \in \mathbb{N}_0^K$, based on the observation that the belief is completely determined by how many times each of the arms has been pulled, $\mathbf{n}_{1:K}$, irrespective of the specific sequence in which the arms have been pulled. Given the pull counts $\mathbf{n}_{1:K}$, we define the payoff of pulling arm a one more time after pulling the individual arms $n_1, \cdots, n_K$ times respectively: with $t = \sum_{a=1}^K n_a$, the effective payoff associated with arm a at time t is

$$r^z(\mathbf{n}_{1:K}, a, \omega) \triangleq \hat{\mu}_{a,n_a}(\omega) - W^{\mathrm{TS}}\left(T-t-1, \mathbf{y}_{t+1}(\mathbf{n}_{1:K}+\mathbf{e}_a, \omega)\right) + W^{\mathrm{TS}}\left(T-t-1, \mathbf{y}_t(\mathbf{n}_{1:K}, \omega)\right), \tag{A.2}$$

where $\mathbf{e}_a \in \mathbb{N}_0^K$ is a basis vector such that the a^{th} component is one and the others are zero. Note that we used the fact that $\mathbb{E}\left[W^{\mathrm{TS}}\left(T-t, \mathbf{y}_t\right) \middle| H_{t-1}\right] = W^{\mathrm{TS}}\left(T-t, \mathbf{y}_{t-1}\right)$.

Consider a subproblem of $(*)$ that maximizes the total payoff given the number of pulls $\mathbf{n}_{1:K}$ across all the arms: with $t = \sum_{a=1}^K n_a$, we get

$$M(\mathbf{n}_{1:K}, \omega) \triangleq \max_{\mathbf{a}_{1:t} \in \mathcal{A}^t} \left\{ \sum_{s=1}^t r_s(\mathbf{a}_{1:s}, \omega) - z_s^{\text{IRS.V-EMAX}}(\mathbf{a}_{1:s}, \omega);\ \sum_{s=1}^t \mathbf{1}\{a_s = a\} = n_a, \forall a \right\}. \tag{A.3}$$

Consequently, the maximal value $M(\mathbf{n}_{1:K}, \omega)$ should satisfy the following Bellman equation:

$$M(\mathbf{n}_{1:K}, \omega) = \max_{a \in \mathcal{A}: n_a \geq 1} \left\{ M(\mathbf{n}_{1:K} - \mathbf{e}_a, \omega) + r^z(\mathbf{n}_{1:K} - \mathbf{e}_a, a, \omega) \right\}, \tag{A.4}$$

i.e., when letting a^* be the maximizer of (A.4), it is optimal to play arm a^* after making the best effort within the allocation $\mathbf{n}_{1:K} - \mathbf{e}_a$. For all feasible counts $\mathbf{n}_{1:K}$'s such that $\sum_{a=1}^K n_a \leq T$, we can compute $M(\mathbf{n}_{1:K}, \omega)$'s by sequentially solving (A.4) in an appropriate order. By doing so, we can obtain the maximal value of the original inner problem $(*)$ by evaluating

$$\max_{\mathbf{n}_{1:K} \in N_T} \left\{ M(\mathbf{n}_{1:K}, \omega) \right\}, \tag{A.5}$$

where $N_T \triangleq \{(n_1, \cdots, n_K) \in \mathbb{N}_0^K : \sum_{a=1}^K n_a = T\}$, and the performance bound $W^{\text{IRS.V-EMAX}}$ is the expected value of (A.5) with respect to the random realization of ω. The optimal action sequence $\mathbf{a}_{1:T}^*$ can be obtained by tracking $M(\mathbf{n}_{1:K}, \omega)$'s backward.

Algorithm 6: Arm selection rule of $\pi^{\text{IRS.V-ZERO}}$ when remaining time is T and current belief is $\mathbf{y}$

Function `IRS.V-EMax`$(T, \mathbf{y})$

$\tilde{\theta}_a \sim \mathcal{P}_a(y_a), \tilde{R}_{a,n} \sim \mathcal{R}_a(\tilde{\theta}), \quad \forall n \in \{1, \ldots, T\}, \forall a \in \{1, \ldots, K\}$

$\tilde{y}_{a,0} \leftarrow y_a, \tilde{y}_{a,n} \leftarrow \mathcal{U}_a(\tilde{y}_{a,n-1}, \tilde{R}_{a,n}), \quad \forall n \in \{1, \ldots, T\}, \forall a \in \{1, \ldots, K\}$

for *each* $\mathbf{n}_{1:K} \in N_{\leq T}$ **do**

  $\tilde{\Gamma}[\mathbf{n}_{1:K}] \leftarrow \mathbb{E}_{\tilde{\mathbf{y}}(\mathbf{n}_{1:K})}\left[\max_a \mu_a(\theta_a)\right]$

end

for *each* $\mathbf{n}_{1:K} \in N_{<T}$ **do**

  $\tilde{r}^z[\mathbf{n}_{1:K}, a] \leftarrow \bar{\mu}_a(\tilde{y}_{a,n_a-1}) + \left(T - \sum_{a=1}^K n_a - 1\right) \times \left(\tilde{\Gamma}[\mathbf{n}_{1:K}] - \tilde{\Gamma}[\mathbf{n}_{1:K} + \mathbf{e}_a]\right), \ \forall a \in \{1, \ldots, K\}$

end

$\tilde{M}[\mathbf{0}] \leftarrow 0$

for *each* $\mathbf{n}_{1:K} \in N_{\leq T} \setminus \{\mathbf{0}\}$ *in order with increasing* $\sum_{a=1}^K n_a$ **do**

  $\tilde{M}[\mathbf{n}_{1:K}] \leftarrow \max_{a:n_a>0}\left\{\tilde{M}[\mathbf{n}_{1:K} - \mathbf{e}_a] + \tilde{r}^z[\mathbf{n}_{1:K} - \mathbf{e}_a, a]\right\}$

  $\tilde{A}[\mathbf{n}_{1:K}] \leftarrow \operatorname{argmax}_{a:n_a>0}\left\{\tilde{M}[\mathbf{n}_{1:K} - \mathbf{e}_a] + \tilde{r}^z[\mathbf{n}_{1:K} - \mathbf{e}_a, a]\right\}$

end

$\mathbf{m}_{1:K} \leftarrow \operatorname{argmax}_{\mathbf{n}_{1:K} \in N_T}\left\{\tilde{M}[\mathbf{n}_{1:K}]\right\}$

for $t = T, \cdots, 1$ **do**

  $\tilde{a}_t^* \leftarrow \tilde{A}[\mathbf{m}_{1:K}]$

  $m_{\tilde{a}_t^*} \leftarrow m_{\tilde{a}_t^*} - 1$

end

return $\tilde{a}_1^*$

Here, $\tilde{\mathbf{y}}(\mathbf{n}_{1:K}) \triangleq (\tilde{y}_{1,n_1}, \cdots, \tilde{y}_{K,n_K})$, $N_{\leq T} \triangleq \{\mathbf{n}_{1:K}; \sum_a n_a \leq T\}$, $N_{<T} \triangleq \{\mathbf{n}_{1:K}; \sum_a n_a < T\}$, and in line 8, $\mathbf{n}_{1:K}$ iterates over $N_{\leq T} \setminus \{\mathbf{0}\}$ in an order in which $\sum_{a=1}^K n_a$ is non-decreasing.

Since $|N_{\leq T}| = O(T^K)$, it requires $O(KT^K)$ operations to compute all $M(\mathbf{n}_{1:K}, \omega)$'s. However, another practical issue is the cost of computing $W^{\mathrm{TS}}(T, \mathbf{y}) = T \times \mathbb{E}_{\mathbf{y}}\left[\max_a \mu_a(\theta_a)\right]$ which has to be evaluated $O(T^K)$ times in total. There is no simple closed-form expression in general, and it should be evaluated with numerical integration or Monte Carlo sampling.

A.2.3 Implementation of IRS.INDEX

We first prove the identity that was utilized in §2.3.5, and then provide the pseudo code for IRS.INDEX policy.

Proposition A.2.1. *The optimization problem* (2.46) *can be reformulated as*

$$\max_{0\leq n\leq T}\left\{T\times\Gamma_0^\lambda+(T-n)\times\left(\lambda-\min_{0\leq i\leq n}\Gamma_i^\lambda\right)+\sum_{i=1}^{n}\left(\hat{\mu}_{a,i-1}-\Gamma_{i-1}^\lambda\right)\right\}. \tag{A.6}$$

Here, the decision variable n is the total number of pulls of a stochastic arm.

Proof. Fix $m \triangleq n_T$, i.e., the total number of pulls on the stochastic arm. Note that if $a_t = 0$, then $(T-t)\times(\Gamma_{n_t}^\lambda - \Gamma_{n_{t-1}}^\lambda) = 0$ since $n_t = n_{t-1}$. The objective function can be represented as

$$\sum_{n=1}^{m}\hat{\mu}_{a,n-1}+(T-m)\times\lambda-\sum_{n=1}^{m}(T-t_n)\times\left(\Gamma_n^\lambda-\Gamma_{n-1}^\lambda\right), \tag{A.7}$$

where $t_n \triangleq \inf\{t; n_t \geq n\}$ represents the time at which the n^{th} pull on the stochastic arm is made. It suffices to find the optimal pulling times $(t_1, \cdots, t_m)$ with $1 \leq t_1 < t_2 < \cdots < t_m \leq T$ by which

$\sum_{n=1}^{m}(T-t_n)\times\left(\Gamma_n^\lambda-\Gamma_{n-1}^\lambda\right)$ is minimized. With $t_0 \triangleq 0$ and $t_{m+1} \triangleq T+1$, we have

$$\sum_{n=1}^{m}(T-t_n)\times\left(\Gamma_n^\lambda-\Gamma_{n-1}^\lambda\right) \tag{A.8}$$

$$=\sum_{n=1}^{m}(T-t_n)\times\Gamma_n^\lambda-\sum_{n=1}^{m}(T-t_n)\times\Gamma_{n-1}^\lambda \tag{A.9}$$

$$=\sum_{n=1}^{m}(T-t_n)\times\Gamma_n^\lambda-\sum_{n=0}^{m-1}(T-t_{n+1})\times\Gamma_n^\lambda \tag{A.10}$$

$$=\sum_{n=0}^{m}(T-t_n)\times\Gamma_n^\lambda-(T-t_0)\times\Gamma_0^\lambda-\sum_{n=0}^{m}(T-t_{n+1})\times\Gamma_n^\lambda+(T-t_{m+1})\times\Gamma_m^\lambda \tag{A.11}$$

$$=-\Gamma_m^\lambda-T\times\Gamma_0^\lambda+\sum_{n=0}^{m}(t_{n+1}-t_n)\times\Gamma_n^\lambda. \tag{A.12}$$

Consider the minimum value among $\Gamma_0^\lambda,\ldots,\Gamma_m^\lambda$ and let $n^* \triangleq \operatorname{argmin}_{0\le n\le m}\Gamma_n^\lambda$. In order to minimize (A.12), it should satisfy that $t_{n+1}-t_n = T-m+1$ for $n=n^*$ and $t_{n+1}-t_n=1$ for $n\neq n^*$. For such t_n's, (A.7) reduces to

$$\sum_{n=1}^{m}\hat{\mu}_{a,n-1}+(T-m)\times\lambda-\left(-\Gamma_m^\lambda-T\times\Gamma_0^\lambda+\sum_{n=0}^{m}\Gamma_n^\lambda+(T-m)\times\min_{0\le n\le m}\Gamma_m^\lambda\right) \tag{A.13}$$

$$=\sum_{n=1}^{m}\hat{\mu}_{a,n-1}+(T-m)\times\left(\lambda-\min_{0\le n\le m}\Gamma_m^\lambda\right)+T\times\Gamma_0^\lambda-\sum_{n=0}^{m-1}\Gamma_n^\lambda. \tag{A.14}$$

By taking its maximum value over $m=0,\cdots,T$, we obtain (A.6). □

The following pseudo code implements the arm selection rule of the IRS.INDEX policy when remaining time is T and current belief is $\mathbf{y}$. In line 14, the infimum can be found via the bisection method, and $\tilde{\mathbf{y}}_{a,0:T} \triangleq (\tilde{y}_{a,0},\ldots,\tilde{y}_{a,T})$ represents the sequence of beliefs under the sampled

outcome.

Algorithm 7: Arm selection rule of IRS.INDEX policy when remaining time is T and current belief is $\mathbf{y}$

Function `IRS.Single.Worth-Trying`$(a, T, \lambda, \tilde{\mathbf{y}}_{a,0:T})$

$\tilde{\Gamma}_n^{\lambda} \leftarrow \mathbb{E}_{\tilde{y}_{a,n}}\left[\max(\mu_a(\theta_a), \lambda)\right], \forall n \in \{0, \ldots, T\}$

$\tilde{S}_{a,0}^{\mu} \leftarrow 0, \tilde{S}_0^{\Gamma} \leftarrow 0, \tilde{m}_0^{\Gamma} \leftarrow \tilde{\Gamma}_0^{\lambda}$

for $n = 1, \cdots, T$ **do**

$\tilde{S}_{a,n}^{\mu} \leftarrow \tilde{S}_{a,n-1}^{\mu} + \bar{\mu}_a(\tilde{y}_{a,n-1})$

$\tilde{S}_n^{\Gamma} \leftarrow \tilde{S}_{a,n-1}^{\Gamma} + \tilde{\Gamma}_n^{\lambda}$

$\tilde{m}_n^{\Gamma} \leftarrow \min\left(\tilde{m}_{n-1}^{\Gamma}, \tilde{\Gamma}_{n-1}^{\lambda}\right)$

end

$\tilde{\varphi}_a \leftarrow \max_{1 \le n \le T}\left\{\tilde{S}_{a,n}^{\mu} + T \times \tilde{\Gamma}_0^{\lambda} + (T-n) \times \left(\lambda - \tilde{m}_n^{\Gamma}\right) - \tilde{S}_n^{\Gamma}\right\} - T \times \lambda$

if $\tilde{\varphi}_a \ge 0$ **then**

return true

else

return false

end

Function `IRS.Index`$(T, \mathbf{y})$

$\tilde{\theta}_a \sim \mathcal{P}_a(y_a), \tilde{R}_{a,n} \sim \mathcal{R}_a(\tilde{\theta}), \quad \forall n \in \{1, \ldots, T\}, \forall a \in \{1, \ldots, K\}$

$\tilde{y}_{a,0} \leftarrow y_a, \quad \tilde{y}_{a,n} \leftarrow \mathcal{U}_a(\tilde{y}_{a,n-1}, \tilde{R}_{a,n}), \quad \forall n \in \{1, \ldots, T\}, \quad \forall a \in \{1, \ldots, K\}$

for $a = 1, \cdots, K$ **do**

$\tilde{\lambda}_a^* \leftarrow \inf\left\{\lambda; \texttt{IRS.Single.Worth-Trying}(a, T, \lambda, \tilde{\mathbf{y}}_{a,0:T}) = \text{true}\right\}$

end

return $\operatorname{argmax}_a \tilde{\lambda}_a^*$

A.3 Proofs for §2.3

Proposition A.3.1 (Mean equivalence)**.** *If the penalty function z_t is dual feasible, the presence of penalties does not affect the performance of a non-anticipating policy π: i.e.,*

$$\mathbb{E}^{\pi}_{\mathbf{y}}\left[\sum_{t=1}^{T} r_t(\mathbf{A}^{\pi}_{1:t},\omega) - z_t(\mathbf{A}^{\pi}_{1:t},\omega)\right] = \mathbb{E}^{\pi}_{\mathbf{y}}\left[\sum_{t=1}^{T} r_t(\mathbf{A}^{\pi}_{1:t},\omega)\right] =: V(\pi,T,\mathbf{y}). \tag{A.15}$$

Proof. The claim immediately follows from the definition of dual feasibility and the linearity of the expectation operator. □

A.3.1 Proof of Theorem 2.3.1

Despite that the results of Theorem 2.3.1 were already well established in [4], we provide the detailed proof as our context is slightly different from that of [4] regarding the measurability of r_t. We define an appending operator $\oplus$ that concatenates an element into a vector so that $\mathbf{a}_{1:t} = \mathbf{a}_{1:t-1} \oplus a_t$.

Weak duality. Define the filtration for the perfect information relaxation $\mathcal{G}_t \triangleq \mathcal{F}_t \cup \sigma(\omega)$ and consider a relaxed policy space $\Pi_{\mathbb{G}} \triangleq \left\{\pi : A^{\pi}_t \text{ is } \mathcal{G}_{t-1}\text{-measurable}, \forall t\right\}$. Then, we have

$$V^*(T,\mathbf{y}) \triangleq \sup_{\pi\in\Pi_{\mathbb{F}}} \mathbb{E}\left[\sum_{t=1}^{T} r_t(\mathbf{A}^{\pi}_{1:t})\right] \overset{\text{Prop A.3.1}}{=} \sup_{\pi\in\Pi_{\mathbb{F}}} \mathbb{E}\left[\sum_{t=1}^{T} r_t(\mathbf{A}^{\pi}_{1:t}) - z_t(\mathbf{A}^{\pi}_{1:t})\right] \tag{A.16}$$

$$\leq \sup_{\pi\in\Pi_{\mathbb{G}}} \mathbb{E}\left[\sum_{t=1}^{T} r_t(\mathbf{A}^{\pi}_{1:t}) - z_t(\mathbf{A}^{\pi}_{1:t})\right] = \mathbb{E}\left[\max_{\mathbf{a}_{1:T}\in\mathcal{A}^T} \sum_{t=1}^{T} r_t(\mathbf{a}_{1:t}) - z_t(\mathbf{a}_{1:t})\right] \tag{A.17}$$

$$= W^z(T,\mathbf{y}), \tag{A.18}$$

where the inequality holds since $\Pi_{\mathbb{F}} \subseteq \Pi_{\mathbb{G}}$. □

Strong duality. Fix T and $\mathbf{y}$. Let $V^{\text{in}}_t(\mathbf{a}_{1:t-1},\omega)$ and $Q^{\text{in}}_t(\mathbf{a}_{1:t-1},a,\omega)$ be, respectively, the value function and the state-action value (Q-value) function that are associated with the inner problem $(*)$ given a particular outcome ω under the ideal penalty (2.22). With $V^{\text{in}}_{T+1} \equiv 0$, we have the

following Bellman equation for the inner problem:

$$Q_t^{\text{in}}(\mathbf{a}_{1:t-1}, a, \omega) \triangleq r_t(\mathbf{a}_{1:t-1} \oplus a, \omega) - z_t^{\text{ideal}}(\mathbf{a}_{1:t-1} \oplus a, \omega) + V_{t+1}^{\text{in}}(\mathbf{a}_{1:t-1} \oplus a, \omega), \tag{A.19}$$

$$V_t^{\text{in}}(\mathbf{a}_{1:t-1}, \omega) = \max_{a \in \mathcal{A}} \left\{ Q_t^{\text{in}}(\mathbf{a}_{1:t-1}, a, \omega) \right\}. \tag{A.20}$$

We argue by induction to show that

$$V_t^{\text{in}}(\mathbf{a}_{1:t-1}, \omega) = V^*(T - t + 1, \mathbf{y}_{t-1}(\mathbf{a}_{1:t-1}, \omega)), \tag{A.21}$$

$$Q_t^{\text{in}}(\mathbf{a}_{1:t-1}, a, \omega) = Q^*(T - t + 1, \mathbf{y}_{t-1}(\mathbf{a}_{1:t-1}, \omega), a), \tag{A.22}$$

for all $\mathbf{a}_{1:t-1} \in \mathcal{A}^{t-1}$, $a \in \mathcal{A}$ and $t \in \{1, \ldots, T+1\}$.

As a terminal case, when $t = T + 1$, the claim holds trivially, since $V_{T+1}^{\text{in}}(\mathbf{a}_{1:T}, \omega) = 0 = V^*(0, \mathbf{y}_T(\mathbf{a}_{1:T}, \omega))$. Now assume that the claim holds for $t + 1$: i.e., $V_{t+1}^{\text{in}}(\mathbf{a}_{1:t}, \omega) = V^*(T - t, \mathbf{y}_t(\mathbf{a}_{1:t}, \omega))$ for all $\mathbf{a}_{1:t} \in \mathcal{A}^t$. For any $\mathbf{a}_{1:t-1} \in \mathcal{A}^{t-1}$ and $a \in \mathcal{A}$, then,

$$Q_t^{\text{in}}(\mathbf{a}_{1:t-1}, a, \omega) = r_t(\mathbf{a}_{1:t-1} \oplus a, \omega) - z_t^{\text{ideal}}(\mathbf{a}_{1:t-1} \oplus a, \omega) + V_{t+1}^{\text{in}}(\mathbf{a}_{1:t-1} \oplus a, \omega) \tag{A.23}$$

$$= \mathbb{E}\left[r_t(\mathbf{a}_{1:t-1} \oplus a, \omega) + V^*\left(T - t, \mathbf{y}_t(\mathbf{a}_{1:t-1} \oplus a, \omega)\right) \middle| H_{t-1}(\mathbf{a}_{1:t-1}, \omega) \right] \tag{A.24}$$

$$\underbrace{-V^*\left(T - t, \mathbf{y}_t(\mathbf{a}_{1:t-1} \oplus a, \omega)\right) + V_{t+1}^{\text{in}}(\mathbf{a}_{1:t-1} \oplus a, \omega)}_{=0} \tag{A.25}$$

$$= \mathbb{E}\left[r_t(\mathbf{a}_{1:t-1} \oplus a, \omega) + V^*\left(T - t, \mathbf{y}_t(\mathbf{a}_{1:t-1} \oplus a, \omega)\right) \middle| H_{t-1}(\mathbf{a}_{1:t-1}, \omega) \right] \tag{A.26}$$

$$= \mathbb{E}_{\mathbf{y}_{t-1}(\mathbf{a}_{1:t-1}, \omega)}\left[R_a + V^*\left(T - t, \mathcal{U}(\mathbf{y}_{t-1}(\mathbf{a}_{1:t-1}, \omega), a, R_a)\right) \right] \tag{A.27}$$

$$= Q^*(T - t, \mathbf{y}_{t-1}(\mathbf{a}_{1:t-1}, \omega), a), \tag{A.28}$$

where the last equality follows from the original Bellman equation (2.15). Consequently, we obtain

$$V_t^{\text{in}}(\mathbf{a}_{1:t-1}, \omega) = \max_{a \in \mathcal{A}} \left\{ Q_t^{\text{in}}(\mathbf{a}_{1:t-1}, a, \omega) \right\} \tag{A.29}$$

$$= \max_{a \in \mathcal{A}} \left\{ Q^*(T - t, \mathbf{y}_{t-1}(\mathbf{a}_{1:t-1}, \omega), a) \right\} \tag{A.30}$$

$$= V^*(T - t, \mathbf{y}_{t-1}(\mathbf{a}_{1:t-1}, \omega)). \tag{A.31}$$

Therefore the claim holds for all $t = 1, \cdots, T$. In particular for $t = 1$, we have

$$V_1^{\text{in}}(\emptyset, \omega) = V^*(T, \mathbf{y}), \quad Q_1^{\text{in}}(\emptyset, a, \omega) = Q^*(T, \mathbf{y}, a), \quad \forall \omega. \tag{A.32}$$

Note that the maximal value of the inner problem does not depend on the outcome ω, i.e., it is deterministic with respect to the randomness of ω. As its expected value, $W^{\text{ideal}}(T, \mathbf{y}) = V^*(T, \mathbf{y})$.

□

A.3.2 Proof of Remark 2.3.1

We proceed on the proof of strong duality. The policy π^{ideal} solves the same inner problem with respect to a randomly sampled outcome $\tilde{\omega}$. When the remaining time is T and the current belief is $\mathbf{y}$, it takes an action with the largest Q-value: together with (A.32), it yields

$$a^{\pi^{\text{ideal}}} = \operatorname*{argmax}_a Q_1^{\text{in}}(\emptyset, a, \tilde{\omega}) = \operatorname*{argmax}_a Q^*(T, \mathbf{y}, a). \tag{A.33}$$

Therefore, at each moment, irrespective of the sampled outcome $\tilde{\omega}$, the policy π^{ideal} always takes the same action that the Bayesian optimal policy would take. Although there might be some ambiguity regarding tie breaking in argmax, it does not affect the expected performance. Therefore, $V(\pi^{\text{ideal}}, T, \mathbf{y}) = V^*(T, \mathbf{y})$. □

A.3.3 Proof of Remark 2.3.2

First observe that for any non-anticipating policy $\pi \in \Pi_{\mathbb{F}}$, since A_t^π is $\mathcal{F}_{t-1}$-measurable, we have

$$\mathbb{E}_{\mathbf{y}}\left[\sum_{t=1}^{T} r_t(\mathbf{A}_{1:t}^\pi, \omega)\right] = \mathbb{E}_{\mathbf{y}}\left[\sum_{t=1}^{T} \mathbb{E}\left(r_t(\mathbf{A}_{1:t}^\pi, \omega)\middle|\mathcal{F}_{t-1}, \boldsymbol{\theta}\right)\right] = \mathbb{E}_{\mathbf{y}}\left[\sum_{t=1}^{T} \mu_{A_t^\pi}(\theta_{A_t^\pi})\right]. \tag{A.34}$$

Since $\mathbb{E}[r_t(\mathbf{a}_{1:t}, \omega)|\boldsymbol{\theta}] = \mu_{a_t}(\theta_{a_t})$ for any $\mathbf{a}_{1:t} \in \mathcal{A}^t$, we further deduce that

$$\mathbb{E}_{\mathbf{y}}\left[\sum_{t=1}^{T} z_t^{\mathrm{TS}}(\mathbf{A}_{1:t}^\pi, \omega)\right] = \mathbb{E}_{\mathbf{y}}\left[\sum_{t=1}^{T} r_t(\mathbf{A}_{1:t}^\pi, \omega)\right] - \mathbb{E}_{\mathbf{y}}\left[\sum_{t=1}^{T} \mu_{A_t^\pi}(\theta_{A_t^\pi})\right] = 0, \tag{A.35}$$

and thus z_t^{TS} is dual feasible.

Also observe that $\mathbb{E}[r_t(\mathbf{a}_{1:t})|\hat{\boldsymbol{\mu}}_{T-1}] = \mathbb{E}[\mu_{a_t}|\hat{\boldsymbol{\mu}}_{T-1}] = \mathbb{E}[\mu_{a_t}|\hat{\boldsymbol{\mu}}_{T-1}, H_{t-1}]$ and $\mathbb{E}[r_t(\mathbf{a}_{1:t})|H_{t-1}] = \mathbb{E}[\mu_{a_t}|H_{t-1}]$ for any $\mathbf{a}_{1:t} \in \mathcal{A}^t$. We can easily verify that each of penalty functions (2.22)–(2.26) has a form of

$$z_t(\mathbf{a}_{1:t}, \omega) = z_t^{\mathrm{TS}}(\mathbf{a}_{1:t}, \omega) + w_t(\mathbf{a}_{1:t}, \omega) - \mathbb{E}[w_t(\mathbf{a}_{1:t}, \omega)|G_{t-1}(\mathbf{a}_{1:t-1}, \omega)], \tag{A.36}$$

for some deterministic function w_t and some relaxed information set $G_{t-1} \supseteq H_{t-1}$. By invoking Proposition 2.3 (iii) of [4], we have that $z_t^{\textsc{Irs.FH}} - z_t^{\mathrm{TS}}$, $z_t^{\textsc{Irs.V-Zero}} - z_t^{\mathrm{TS}}$, $z_t^{\textsc{Irs.V-EMax}} - z_t^{\mathrm{TS}}$, and $z_t^{\mathrm{ideal}} - z_t^{\mathrm{TS}}$ are dual feasible, and therefore so are $z_t^{\textsc{Irs.FH}}$, $z_t^{\textsc{Irs.V-Zero}}$, $z_t^{\textsc{Irs.V-EMax}}$, and z_t^{ideal}. □

A.4 Proofs for §2.4

A.4.1 Notes on regularity

Proposition A.4.1. *If* $\mathbb{E}_{\mathbf{y}}|R_{a,n}| < \infty$ *for all* a*,*

$$\mathbb{E}_{\mathbf{y}}|\mu_a(\theta_a)| < \infty, \quad \textit{and} \quad W^{\mathrm{TS}}(T, \mathbf{y}) < \infty, \quad \forall T \in \mathbb{N}. \tag{A.37}$$

Proof. By Jensen's inequality,

$$\mathbb{E}_{\mathbf{y}}|\mu_a(\theta_a)| = \mathbb{E}_{\mathbf{y}}\left[\left|\mathbb{E}\left(R_{a,n}|\theta_a\right)\right|\right] \leq \mathbb{E}_{\mathbf{y}}\left[\mathbb{E}\left(\left|R_{a,n}\right|\middle|\theta_a\right)\right] = \mathbb{E}_{\mathbf{y}}|R_{a,n}| < \infty. \tag{A.38}$$

Consequently,

$$\mathbb{E}_{\mathbf{y}}\left[\max_a \mu_a(\theta_a)\right] \leq \mathbb{E}_{\mathbf{y}}\left[\sum_{a=1}^{K}|\mu_a(\theta_a)|\right] = \sum_{a=1}^{K}\mathbb{E}_{\mathbf{y}}|\mu_a(\theta_a)| < \infty. \tag{A.39}$$

The claim holds since $W^{\mathrm{TS}}(T, \mathbf{y}) = T \times \mathbb{E}_{\mathbf{y}}[\max_a \mu_a(\theta_a)]$. □

Proposition A.4.2. *If* $\mathbb{E}_{\mathbf{y}}|R_{a,n}| < \infty$*,*

$$\lim_{n\to\infty}\hat{\mu}_{a,n}(\omega; y_a) = \lim_{n\to\infty}\frac{1}{n}\sum_{i=1}^{n}R_{a,i} = \mu_a(\theta_a) \quad \textit{almost surely}, \tag{A.40}$$

where $\hat{\mu}_{a,n}(\omega; y_a) \triangleq \mathbb{E}_{y_a}\left[\mu_a(\theta_a)|R_{a,1}, \cdots, R_{a,n}\right]$.

Proof. Fix a and let $\mathcal{H}_n \triangleq \sigma\left(R_{a,1}, \cdots, R_{a,n}\right)$. First note that, by the strong law of large numbers, $\lim_{n\to\infty}\frac{1}{n}\sum_{i=1}^{n}R_{a,i} = \mu_a(\theta_a)$ almost surely. Therefore, $\mu_a(\theta_a)$ is measurable with respect to $\mathcal{H}_\infty \triangleq \bigcup_n \mathcal{H}_n$. Also note that $\hat{\mu}_{a,n} = \mathbb{E}\left(\mu_a(\theta_a)|\mathcal{H}_n\right)$ is a Doob martingale adapted to $\mathcal{H}_n$. By Levy's upward theorem, since $\mu_a(\theta_a) \in \mathcal{L}^1$ by Proposition A.4.1, $\hat{\mu}_{a,n}$ converges to $\mathbb{E}\left(\mu_a(\theta_a)|\mathcal{H}_\infty\right) = \mu_a(\theta_a)$ almost surely as $n \to \infty$. □

A.4.2 Proof of Proposition 2.4.1

Asymptotic behavior of $\pi^{\mathrm{IRS.FH}}$. Let $\tilde{\omega}$ be the sampled outcome used by $\pi^{\mathrm{IRS.FH}}$. By Proposition A.4.2, we have $\lim_{n\to\infty}\hat{\mu}_{a,n}(\tilde{\omega}) = \mu_a(\tilde{\theta}_a)$ for almost all $\tilde{\omega}$. This, together with the assumption that $\mu_i(\theta_i) \neq \mu_j(\theta_j)$ for $i \neq j$, since $\operatorname{argmax}_a \mu_a(\tilde{\theta}_a)$ is uniquely defined for almost all $\tilde{\omega}$, yields

$$\operatorname*{argmax}_a \mu_a(\tilde{\theta}_a) = \operatorname*{argmax}_a \lim_{n\to\infty}\hat{\mu}_{a,n}(\tilde{\omega}) = \lim_{n\to\infty}\operatorname*{argmax}_a \hat{\mu}_{a,n}(\tilde{\omega}) \quad \text{a.s.} \tag{A.41}$$

Since almost-sure convergence guarantees convergence in distribution, for any $a \in \mathcal{A}$,

$$\lim_{T\to\infty} \mathbb{P}\left[A^{\text{IRS.FH}}(T, \mathbf{y}) = a\right] = \lim_{T\to\infty} \mathbb{P}\left[\operatorname*{argmax}_{a'} \hat{\mu}_{a',T-1}(\tilde{\omega}) = a\right] \tag{A.42}$$

$$= \mathbb{P}\left[\operatorname*{argmax}_{a'} \mu_{a'}(\tilde{\theta}_{a'}) = a\right] \tag{A.43}$$

$$= \mathbb{P}\left[A^{\text{TS}}(\mathbf{y}) = a\right]. \tag{A.44}$$

Note that we are not assuming that $\pi^{\text{IRS.FH}}$ and π^{TS} share the randomness. The sampled parameters used in π^{TS} are not necessarily the ones used in $\pi^{\text{IRS.FH}}$, but their distributions are identical since they are drawn from the same prior. □

Asymptotic behavior of $\pi^{\text{IRS.V-ZERO}}$. To simplify notation, let $A^{\circ}_T \triangleq A^{\text{IRS.V-ZERO}}(T, \mathbf{y})$. As above, it suffices to show that $\lim_{T\to\infty} A^{\circ}_T = \operatorname{argmax}_{a\in\mathcal{A}} \mu_a(\tilde{\theta}_a) := A^{\text{TS}}$ for almost all sampled outcome $\tilde{\omega}$. We hide $\tilde{\omega}$ and $\tilde{\theta}_a$ from the notation for the further simplication.

Define

$$\Delta \triangleq \min_{a\neq A^{\text{TS}}} |\mu_{A^{\text{TS}}} - \mu_a| \quad \text{and} \quad M \triangleq \sup_{a\in\mathcal{A},n\geq 0} \left|\hat{\mu}_{a,n}\right|. \tag{A.45}$$

We have $0 < \Delta < 2M < \infty$ almost surely since $\mu_i(\tilde{\theta}_i) \neq \mu_j(\tilde{\theta}_j)$ for $i \neq j$ and $\lim_{n\to\infty} \hat{\mu}_{a,n} = \mu_a < \infty$ almost surely for all a. In addition, there exists $N \in \mathbb{N}$ such that

$$\left|\hat{\mu}_{a,n} - \mu_a\right| < \frac{\Delta}{4}, \quad \forall n \geq N, \quad \forall a \in \mathcal{A}. \tag{A.46}$$

For such N, we have

$$\inf_{n\geq N} \hat{\mu}_{A^{\text{TS}},n} \geq \sup_{n\geq N} \hat{\mu}_{a,n} + \frac{\Delta}{2}, \quad \forall a \neq A^{\text{TS}}. \tag{A.47}$$

Note that A^{TS}, Δ, M, and N do not have the dependency on T.

To argue by contradiction, suppose that $A^{\circ}_T \neq A^{\text{TS}}$ for some large T such that $T \geq 2N + \frac{8MN}{\Delta} + 2$. Define the optimal allocation to the inner problem of IRS.V-ZERO for such T:

$$\mathbf{n}^{\circ}_{1:K} \triangleq \operatorname*{argmax}_{\mathbf{n}_{1:K}\in N_T} \left\{\sum_{a=1}^{K} \sum_{s=1}^{n_a} \hat{\mu}_{a,s-1}\right\}, \tag{A.48}$$

where the ties are broken arbitrarily in argmax{}. We let $n^\circ(a)$ be the a^{th} component of $\mathbf{n}^\circ_{1:K}$. According to the specified arm selection rule, we have $A^\circ_T = \text{argmax}_a\, n^\circ(a)$ and hence $n^\circ(A^\circ_T) \geq \lfloor \frac{T}{2} \rfloor$ ($> N$). We prove the claim for the following two cases:

Case 1: If $n^\circ(a^{\text{TS}}) \geq N$, consider an allocation $\mathbf{n}^\dagger_{1:K}$ that is a deviation from the given optimal allocation $\mathbf{n}^\circ_{1:K}$ such that arm a^{TS} gets one pull whereas arm A°_T gets one less pull: i.e., $n^\dagger(A^{\text{TS}}) = n^\circ(A^{\text{TS}}) + 1$, $n^\dagger(A^\circ_T) = n^\circ(A^\circ_T) - 1$, and $n^\dagger(a) = n^\circ(a)$ for any $a \notin \{A^{\text{TS}}, A^\circ_T\}$. The change in the total payoff from this deviation is

$$\sum_{a=1}^{K} \sum_{i=1}^{n^\dagger(a)} \hat{\mu}_{a,i-1} - \sum_{a=1}^{K} \sum_{i=1}^{n^\circ(a)} \hat{\mu}_{a,i-1} = \hat{\mu}_{A^{\text{TS}},n^\circ(A^{\text{TS}})} - \hat{\mu}_{A^\circ_T,n^\circ(A^\circ_T)-1} \geq \frac{\Delta}{2} > 0, \tag{A.49}$$

where the inequality follows from (A.47) and that $n^\circ(A^{\text{TS}}) \geq N$ and $n^\circ(A^\circ_T) \geq N$. The allocation $\mathbf{n}^\dagger_{1:K}$ is strictly better than $\mathbf{n}^\circ_{1:K}$, which contradicts the assumption that $\mathbf{n}^\circ_{1:K}$ is an optimal allocation.

Case 2: If $n^\circ(A^{\text{TS}}) < N$, consider an allocation $\mathbf{n}^\dagger_{1:K}$ that is a deviation from the given optimal allocation $\mathbf{n}^\circ_{1:K}$ such that arm A°_T gets no more than N pulls whereas arm A^{TS} gets the remains: i.e.,

$$n^\dagger(a) \triangleq \begin{cases} n^\circ(A^{\text{TS}}) + (n^\circ(A^\circ_T) - N) & \text{if } a = A^{\text{TS}}, \\ N & \text{if } a = A^\circ_T, \\ n^\circ(a) & \text{if } a \notin \{A^{\text{TS}}, A^\circ_T\}. \end{cases} \tag{A.50}$$

By making this the deviation, the total payoff should increase by

$$\sum_{a=1}^{K}\sum_{i=1}^{n^{\dagger}(a)}\hat{\mu}_{a,i-1}-\sum_{a=1}^{K}\sum_{i=1}^{n^{\circ}(a)}\hat{\mu}_{a,i-1} \tag{A.51}$$

$$=\sum_{i=n^{\circ}(A^{\mathrm{TS}})+1}^{n^{\circ}(A^{\mathrm{TS}})+(n^{\circ}(A_T^{\circ})-N)}\hat{\mu}_{A^{\mathrm{TS}},i-1}-\sum_{i=N+1}^{n^{\circ}(A_T^{\circ})}\hat{\mu}_{A_T^{\circ},i-1} \tag{A.52}$$

$$\geq -(N-n^{\circ}(A^{\mathrm{TS}}))\cdot 2M+\sum_{i=N+1}^{n^{\circ}(A_T^{\circ})}\hat{\mu}_{A^{\mathrm{TS}},i-1}-\sum_{i=N+1}^{n^{\circ}(A_T^{\circ})}\hat{\mu}_{A_T^{\circ},i-1} \tag{A.53}$$

$$\geq -(N-n^{\circ}(A^{\mathrm{TS}}))\cdot 2M+(n^{\circ}(A_T^{\circ})-N)\cdot\frac{\Delta}{2} \tag{A.54}$$

$$\geq (n^{\circ}(A_T^{\circ})-N)\cdot\frac{\Delta}{2}-2NM. \tag{A.55}$$

Since $T \geq 2N+\frac{8MN}{\Delta}+2$ and $n^{\circ}(A_T^{\circ}) \geq \lfloor\frac{T}{2}\rfloor$, the last term is strictly positive, which is a contradiction.

We've shown that for almost all $\tilde{\omega}$, when T is large enough, the optimal allocation $\mathbf{n}^{\circ}_{1:K}$ must allocate more than a half of the pulls on arm $A^{\mathrm{TS}} = \mathrm{argmax}_a\, \mu_a(\tilde{\theta}_a)$. This concludes the proof.

A.4.3 Proof of Theorem 2.4.1

Proof of "$W^{\mathrm{TS}}(T,\mathbf{y}) \geq W^{\textsc{Irs.FH}}(T,\mathbf{y})$"

Proof. It immediately follows from Jensen's inequality: since $\max(\cdots)$ is a convex function,

$$W^{\mathrm{TS}}(T,\mathbf{y}) = T\times\mathbb{E}_{\mathbf{y}}\left[\max_a \mu_a(\theta_a)\right] \geq T\times\mathbb{E}_{\mathbf{y}}\left[\max_a \mathbb{E}\left(\mu_a(\theta_a)|\,\hat{\boldsymbol{\mu}}_{T-1}\right)\right] = W^{\textsc{Irs.FH}}(T,\mathbf{y}). \tag{A.56}$$

□

Proof of "$W^{\textsc{Irs.FH}}(T,\mathbf{y}) \geq W^{\textsc{Irs.V-Zero}}(T,\mathbf{y})$"

Lemma A.4.1 (Variant of Jensen's inequality)**.** *Suppose that $\varphi : \mathbb{R} \mapsto \mathbb{R}$ is an **increasing** (deterministic) function. Then, for any real-valued random variable X such that $\mathbb{E}|X| < \infty$,*

$$\mathbb{E}\left[\max\left\{X+\varphi(X),0\right\}\right] \geq \mathbb{E}\left[\max\left\{\mathbb{E}(X)+\varphi(X),0\right\}\right]. \tag{A.57}$$

Proof. Define $\mu \triangleq \mathbb{E}(X)$ and $f_x(t) \triangleq \max\{t + \varphi(x), 0\}$. Since $f_x(\cdot)$ is a convex function for each $x \in \mathbb{R}$,

$$f_x(t) \geq f_x(\mu) + (t - \mu) \cdot f_x'(\mu) = \max\{\mu + \varphi(x), 0\} + (t - \mu) \cdot \mathbf{1}\{\mu + \varphi(x) \geq 0\}, \quad \forall t, \quad \forall x. \tag{A.58}$$

By setting $t = x$, we get

$$\max\{x + \varphi(x), 0\} = f_x(x) \geq \max\{\mu + \varphi(x), 0\} + (x - \mu) \cdot \mathbf{1}\{\mu + \varphi(x) \geq 0\}, \quad \forall x. \tag{A.59}$$

Note that, since $\mathbf{1}\{\mu + \varphi(x) \geq 0\}$ is increasing in x, (i) for any $x \geq \mu$, $(x - \mu) \geq 0$ and $\mathbf{1}\{\mu + \varphi(x)\} \geq \mathbf{1}\{\mu + \varphi(\mu)\}$, and (ii) for any $x < \mu$, $(x - \mu) < 0$ and $\mathbf{1}\{\mu + \varphi(x)\} \leq \mathbf{1}\{\mu + \varphi(\mu)\}$. Therefore,

$$(x - \mu) \cdot \mathbf{1}\{\mu + \varphi(x) \geq 0\} \geq (x - \mu) \cdot \mathbf{1}\{\mu + \varphi(\mu) \geq 0\}, \quad \forall x \in \mathbb{R}. \tag{A.60}$$

Combining this with (A.59), we get

$$\max\{x + \varphi(x), 0\} \geq \max\{\mu + \varphi(x), 0\} + (x - \mu) \cdot \mathbf{1}\{\mu + \varphi(\mu) \geq 0\}, \quad \forall x \in \mathbb{R}. \tag{A.61}$$

For random variable X, by taking expectation, we get

$$\begin{aligned}
\mathbb{E}\left[\max\{X + \varphi(X), 0\}\right] &\geq \mathbb{E}\left[\max\{\mu + \varphi(X), 0\} + (X - \mu) \cdot \mathbf{1}\{\mu + \varphi(\mu) \geq 0\}\right] && \text{(A.62)} \\
&\geq \mathbb{E}\left[\max\{\mu + \varphi(X), 0\}\right] + \mathbb{E}(X - \mu) \cdot \mathbf{1}\{\mu + \varphi(\mu) \geq 0\} && \text{(A.63)} \\
&= \mathbb{E}\left[\max\{\mu + \varphi(X), 0\}\right]. && \text{(A.64)}
\end{aligned}$$

□

Corollary A.4.1. *On a probability space $(\Omega, \mathcal{F}, \mathbb{P})$, let $\varphi(x, \omega) : \mathbb{R} \times \Omega \mapsto \mathbb{R}$ be a function such that (i) the mapping $x \mapsto \varphi(x, \omega)$ is **increasing** for each $\omega \in \Omega$ and (ii) for some sub-σ-field $\mathcal{H} \subseteq \mathcal{F}$, the mapping $\omega \mapsto \varphi(x, \omega)$ is $\mathcal{H}$-**measurable** for each $x \in \mathbb{R}$ (i.e., $\varphi(\cdot, \omega)$ is a deterministic function conditioned on $\mathcal{H}$). Then*

$$\mathbb{E}\left[\max\left\{X(\omega) + \varphi(X(\omega), \omega), 0\right\}\right] \geq \mathbb{E}\left[\max\left\{\mathbb{E}(X|\mathcal{H})(\omega) + \varphi(X(\omega), \omega), 0\right\}\right]. \tag{A.65}$$

Proof. Define

$$\mu(\omega) \triangleq \mathbb{E}(X|\mathcal{H})(\omega), \quad I(\omega) \triangleq \mathbf{1}\{\mu(\omega) + \varphi(\mu(\omega), \omega) \geq 0\}. \tag{A.66}$$

By (A.61), we have

$$\max\{x + \varphi(x, \omega), 0\} \geq \max\{\mu(\omega) + \varphi(x, \omega), 0\} + (x - \mu(\omega)) \cdot I(\omega), \quad \forall x \in \mathbb{R}, \quad \text{for each } \omega \in \Omega. \tag{A.67}$$

Since $\mu(\omega)$ and $I(\omega)$ are $\mathcal{H}$-measurable,

$$\begin{aligned}
\mathbb{E}\left[\max\{X(\omega) + \varphi(X(\omega), \omega), 0\}\right] &\geq \mathbb{E}\left[\max\{\mu(\omega) + \varphi(X(\omega), \omega), 0\} + (X(\omega) - \mu(\omega)) \cdot I(\omega)\right] && \text{(A.68)} \\
&= \mathbb{E}\left[\mathbb{E}\left(\max\{\mu(\omega) + \varphi(X(\omega), \omega), 0\} + (X(\omega) - \mu(\omega)) \cdot I(\omega) | \mathcal{H}\right)\right] && \text{(A.69)} \\
&= \mathbb{E}\left[\max\{\mu(\omega) + \varphi(X(\omega), \omega), 0\}\right] + \mathbb{E}\left[\mathbb{E}\left(\left(X(\omega) - \mu(\omega)\right) \cdot I(\omega) | \mathcal{H}\right)\right] && \text{(A.70)} \\
&= \mathbb{E}\left[\max\{\mathbb{E}\left(X|\mathcal{H}\right)(\omega) + \varphi(X(\omega), \omega), 0\}\right] && \text{(A.71)} \\
&\quad + \mathbb{E}\left[\underbrace{\left(\mathbb{E}(X|\mathcal{H})(\omega) - \mu(\omega)\right)}_{=0} \cdot I(\omega)\right] && \text{(A.72)} \\
&= \mathbb{E}\left[\max\{\mathbb{E}\left(X|\mathcal{H}\right)(\omega) + \varphi(X(\omega), \omega), 0\}\right]. && \text{(A.73)}
\end{aligned}$$

□

Corollary A.4.2. *On a probability space $(\Omega, \mathcal{F}, \mathbb{P})$, let $(C_0, \cdots, C_T)$ be* ***$\mathcal{H}$-measurable real-valued random variables*** *for some sub-σ-field $\mathcal{H} \subseteq \mathcal{F}$ (i.e., C_i's are constants conditioned on $\mathcal{H}$). Then*

$$\mathbb{E}\left[\max_{0 \leq i \leq T}\left\{(i-n)^+ \times X + C_i\right\}\right] \geq \mathbb{E}\left[\max_{0 \leq i \leq T}\left\{\mathbb{E}\left(X|\mathcal{H}\right) \cdot \mathbf{1}\{i \geq n+1\} + (i-n-1)^+ \times X + C_i\right\}\right] \tag{A.74}$$

for any $n = 0, 1, \cdots, T$.

Proof. When $n = T$, both sides become $\mathbb{E}\left[\max_{0 \leq i \leq T}\{C_i\}\right]$, which makes the claim true. Fix $n < T$ and

define

$$\varphi(x,\omega) \triangleq \max_{n+1\le i\le T}\left\{(i-n-1)\times x + C_i(\omega)\right\} - \max_{0\le i\le n}\left\{C_i(\omega)\right\}. \tag{A.75}$$

Note that $\varphi(x,\omega)$ satisfies the conditions in Corollary A.4.1. By Corollary A.4.1,

$$\mathbb{E}\left[\max_{0\le i\le T}\left\{(i-n)^+\times X + C_i\right\}\right] \tag{A.76}$$

$$= \mathbb{E}\left[\max\left\{\max_{n+1\le i\le T}\left\{(i-n)\times X + C_i\right\},\quad \max_{0\le i\le n} C_i\right\}\right] \tag{A.77}$$

$$= \mathbb{E}\left[\max\left\{X + \max_{n+1\le i\le T}\left\{(i-n-1)\times X + C_i\right\},\quad \max_{0\le i\le n} C_i\right\}\right] \tag{A.78}$$

$$= \mathbb{E}\left[\max\left\{X(\omega) + \underbrace{\max_{n+1\le i\le T}\left\{(i-n-1)\times X(\omega) + C_i(\omega)\right\} - \max_{0\le i\le n} C_i(\omega)}_{=\varphi(X(\omega),\omega)},\quad 0\right\} + \max_{0\le i\le n} C_i(\omega)\right] \tag{A.79}$$

$$\ge \mathbb{E}\left[\max\left\{\mathbb{E}\left(X|\mathcal{H}\right)(\omega) + \max_{n+1\le i\le T}\left\{(i-n-1)\times X(\omega) + C_i(\omega)\right\} - \max_{0\le i\le n} C_i(\omega),\quad 0\right\} + \max_{0\le i\le n} C_i(\omega)\right] \tag{A.80}$$

$$= \mathbb{E}\left[\max\left\{\max_{n+1\le i\le T}\left\{\mathbb{E}\left(X|\mathcal{H}\right) + (i-n-1)\times X + C_i\right\},\quad \max_{0\le i\le n} C_i\right\}\right] \tag{A.81}$$

$$= \mathbb{E}\left[\max_{0\le i\le T}\left\{\mathbb{E}\left(X|\,\mathcal{H}\right)\cdot\mathbf{1}\{i\ge n+1\} + (i-n-1)^+\times X + C_i\right\}\right]. \tag{A.82}$$

□

Proof of "$W^{\text{IRS.FH}}(T,\mathbf{y}) \ge W^{\text{IRS.V-ZERO}}(T,\mathbf{y})$." Define

$$N_T \triangleq \left\{\mathbf{n}_{1:K}\in\mathbb{N}_0^K : \sum_{a=1}^K n_a = T\right\} \quad\text{and}\quad S_a(n_a) \triangleq \sum_{i=1}^{n_a}\hat{\mu}_{a,i-1}. \tag{A.83}$$

What we want to show is

$$W^{\text{IRS.FH}} \equiv \mathbb{E}\left[T\times\max_a\{\hat{\mu}_{a,T-1}\}\right] = \mathbb{E}\left[\max_{\mathbf{n}_{1:K}\in N_T}\left\{\sum_{a=1}^K n_a\times\hat{\mu}_{a,T-1}\right\}\right] \tag{A.84}$$

$$\ge \mathbb{E}\left[\max_{\mathbf{n}_{1:K}\in N_T}\left\{\sum_{a=1}^K S_a(n_a)\right\}\right] \equiv W^{\text{IRS.V-ZERO}}. \tag{A.85}$$

Further define

$$U_{k,n} \triangleq \mathbb{E}\left[\max_{\mathbf{n}_{1:K}\in N_T}\left\{\left(\sum_{a=1}^{k-1} S_a(n_a)\right) + \left(S_k(n_k \wedge n) + (n_k - n)^+ \times \hat{\mu}_{a,T-1}\right) + \left(\sum_{a=k+1}^{K} n_a \times \hat{\mu}_{a,T-1}\right)\right\}\right], \quad \text{(A.86)}$$

where $a \wedge b \triangleq \min(a, b)$. Observe that $W^{\text{IRS.FH}} = U_{1,0}$, $W^{\text{IRS.V-ZERO}} = U_{K,T}$, and $U_{k+1,0} = U_{k,T}$. Therefore, it suffices to show that

$$U_{k,n} \geq U_{k,n+1}, \quad \forall k = 1, \cdots, K, \quad \forall n = 0, \cdots, T-1. \quad \text{(A.87)}$$

Fix k and n. Define a sub-σ-field

$$\mathcal{H} \triangleq \sigma\left(\{R_{a,s}\}_{a=k,1\leq s\leq n} \cup \{R_{a,s}\}_{a\neq k,1\leq s\leq T-1}\right). \quad \text{(A.88)}$$

For each $i = 0, \cdots, T$, define

$$C_i \triangleq \max\left\{\left(\sum_{a=1}^{k-1} S_a(n_a)\right) + S_k(i \wedge n) + \left(\sum_{a=k+1}^{K} n_a \times \hat{\mu}_{a,T-1}\right) \;:\; \mathbf{n}_{1:K} \in N_T, n_k = i\right\}. \quad \text{(A.89)}$$

Note that C_i's are $\mathcal{H}$-measurable and

$$U_{k,n} = \mathbb{E}\left[\max_{0\leq i\leq T}\left\{(i-n)^+ \times \hat{\mu}_{k,T-1} + C_i\right\}\right]. \quad \text{(A.90)}$$

With $X \triangleq \hat{\mu}_{a,T-1}$,

$$\begin{aligned}
U_{k,n} &= \mathbb{E}\left[\max_{0\leq i\leq T}\left\{(i-n)^+ \times X + C_i\right\}\right] && \text{(A.91)}\\
&\overset{\text{Corollary A.4.2}}{\geq} \mathbb{E}\left[\max_{0\leq i\leq T}\left\{\mathbb{E}(X|\mathcal{H})\cdot\mathbf{1}\{i\geq n+1\} + (i-n-1)^+ \times X + C_i\right\}\right] && \text{(A.92)}\\
&\overset{\text{(a)}}{=} \mathbb{E}\left[\max_{0\leq i\leq T}\left\{\hat{\mu}_{k,n}\cdot\mathbf{1}\{i\geq n+1\} + (i-n-1)^+ \times \hat{\mu}_{a,T-1} + C_i\right\}\right] && \text{(A.93)}\\
&\overset{\text{(b)}}{=} U_{k,n+1}. && \text{(A.94)}
\end{aligned}$$

Equation (a) holds since $\mathbb{E}(X|\mathcal{H}) = \mathbb{E}\left(\hat{\mu}_{k,T-1}|\mathcal{H}\right) = \mathbb{E}\left(\hat{\mu}_{k,T-1}|R_{k,1}, \cdots, R_{k,n}\right) = \hat{\mu}_{a,n}$, and equation (b) holds since $S_k(i\wedge n) + \hat{\mu}_{k,n}\cdot\mathbf{1}\{i\geq n+1\} = \sum_{s=1}^{n}\hat{\mu}_{k,s-1}\cdot\mathbf{1}\{i\geq s\} + \hat{\mu}_{k,n}\cdot\mathbf{1}\{i\geq n+1\} = \sum_{s=1}^{n+1}\hat{\mu}_{k,s-1}\cdot\mathbf{1}\{i\geq$

$s\} = S_k(i \wedge (n+1))$. □

A note on the proof. One may wonder if the above result can be derived in a simpler way by exploiting the properties of nested filtration [e.g., Proposition 2.3 of 4]. Unlike the proof of $W^{\text{TS}} \geq W^{\text{IRS.FH}}$, however, the proof of $W^{\text{IRS.FH}} \geq W^{\text{IRS.V-ZERO}}$ does not simply follow from the fact that $\sigma(\hat{\boldsymbol{\mu}}_{T-1})$ is larger than $\sigma(H_{t-1})$.

Consider a Bernoulli MAB with $K = 2$, $T = 2$, and a prior distribution Beta$(1, 1)$, and let us introduce its variation whose reward function is given by $r'_t(\cdot)$ as follows:

$$r'_1(a_1) = r_1(a_1), \quad r'_2(\mathbf{a}_{1:2}) = -\kappa r_2(\mathbf{a}_{1:2}), \tag{A.95}$$

where $r_t(\cdot)$ is the reward function of the original Bernoulli MAB. When $\kappa > 0$, one can show that

$$W^{\text{IRS.FH}} = \mathbb{E}\left[\max_{\mathbf{a}_{1:T}} \left\{\sum_{t=1}^{T} \mathbb{E}(r'_t(\mathbf{a}_{1:t})|\hat{\boldsymbol{\mu}}_{T-1})\right\}\right] = \frac{7}{12} - \frac{5}{12}\kappa, \tag{A.96}$$

$$W^{\text{IRS.V-ZERO}} = \mathbb{E}\left[\max_{\mathbf{a}_{1:T}} \left\{\sum_{t=1}^{T} \mathbb{E}(r'_t(\mathbf{a}_{1:t})|H_{t-1})\right\}\right] = \frac{1}{2} - \frac{3}{8}\kappa. \tag{A.97}$$

If κ is large enough, we obtain $W^{\text{IRS.FH}} < W^{\text{IRS.V-ZERO}}$, which is opposite to the above result.

Recall that the additional gain from knowing the future information can be decomposed into two components; the gain from knowing the immediate reward and the gain from knowing the next belief state, where IRS.V-ZERO considers the former component only. When those two components are not aligned as in this example (i.e., a higher r'_1 leads to a worse next belief state), the DM can exploit the penalties if they penalize only for the first component (e.g., when r'_1 is smaller than expected, the DM will get compensated for this difference but she can still earn the larger reward in the next period).

This is also related to the fact that $z_t^{\text{IRS.V-ZERO}}$ does not correspond to zero penalty under the some (partial) information relaxation, but should be understood as an approximation of z_t^{ideal} under the perfect information relaxation. As opposed to TS and IRS.FH, the optimal solution to the

IRS.V-ZERO's inner problem may depend on the entire outcome ω. With the terminology of [4], there is a mismatch between the filtration that generates the penalties and the filtration that characterizes the relaxed policy space.

Proof of "$W^{\text{TS}}(T, \mathbf{y}) \geq W^{\text{IRS.V-EMAX}}(T, \mathbf{y})$"

To show that $W^{\text{TS}} \geq W^{\text{IRS.V-EMAX}}$, we take a completely different approach that utilizes Theorem 4 in [10]. We here rephrase the definition and the theorem therein using our notation.

Definition A.4.1 (Supersolution)**.** *An approximate value function $\widehat{V} : \mathbb{N}_0 \times \mathcal{Y} \mapsto \mathbb{R}$ is a **supersolution** to the Bellman equation* (2.15) *if*

$$\widehat{V}(T, \mathbf{y}) \geq \max_{a \in \mathcal{A}} \left\{ \mathbb{E}_{y_a} \left[R_{a,1} + \widehat{V}(T-1, \mathcal{U}(\mathbf{y}, R_{a,1}, r)) \right] \right\}, \quad \forall \mathbf{y} \in \mathcal{Y}, \quad \forall T \geq 1, \tag{A.98}$$

with $\widehat{V}(0, \mathbf{y}) = 0$ for all $\mathbf{y} \in \mathcal{Y}$.

Remark A.4.1. *If $\widehat{V}(\cdot, \cdot)$ is a supersolution, then for any given ω, T, and $\mathbf{y}$,*

$$\widehat{V}(T-t+1, \mathbf{y}_{t-1}(\mathbf{a}_{1:t-1}, \omega; \mathbf{y})) \geq \mathbb{E}_{\mathbf{y}} \left[r_t(\mathbf{a}_{1:t-1} \oplus a, \omega; \mathbf{y}) + \widehat{V}(T-t, \mathbf{y}_t(\mathbf{a}_{1:t-1} \oplus a, \omega; \mathbf{y})) \middle| H_{t-1}(\mathbf{a}_{1:t-1}, \omega) \right], \tag{A.99}$$

for all $a \in \mathcal{A}$, $\mathbf{a}_{1:t-1} \in \mathcal{A}^{t-1}$ and $t \in \{1, \ldots, T\}$.

Lemma A.4.2 (Theorem 4 of [10], rephrased)**.** *Consider a penalty function $\hat{z}_t$ generated by $\widehat{V}(\cdot, \cdot)$:*

$$\begin{aligned} \hat{z}_t(\mathbf{a}_{1:t}, \omega; T, \mathbf{y}) \triangleq\ & r_t(\mathbf{a}_{1:t}, \omega) - \mathbb{E}_{\mathbf{y}} \left[r_t(\mathbf{a}_{1:t}, \omega) \,|H_{t-1}(\mathbf{a}_{1:t-1}, \omega) \right] \\ & + \widehat{V}\left(T-t, \mathbf{y}_t(\mathbf{a}_{1:t}, \omega; \mathbf{y})\right) - \mathbb{E}_{\mathbf{y}} \left[\widehat{V}\left(T-t, \mathbf{y}_t(\mathbf{a}_{1:t}, \omega; \mathbf{y})\right) \middle| H_{t-1}(\mathbf{a}_{1:t-1}, \omega) \right]. \end{aligned} \tag{A.100}$$

If $\widehat{V}(\cdot, \cdot)$ is a supersolution, then the performance bound induced by penalty function $\hat{z}_t$ is tighter than $\widehat{V}$: i.e.,

$$W^{\hat{z}}(T, \mathbf{y}) \leq \widehat{V}(T, \mathbf{y}). \tag{A.101}$$

And this holds in a stronger sense: for each outcome ω, the maximal value of the inner problem with respect to ω (denoted by $V_1^{\hat{z},\mathrm{in}}(\emptyset, \omega; T, \mathbf{y})$ in the proof) is smaller than or equal to $\widehat{V}(T, \mathbf{y})$.

Proof. Let $V_t^{\hat{z},\mathrm{in}}(\cdot)$ be the DP solution of inner problem $(*)$ for a given penalty $\hat{z}_t$ with respect to a particular outcome ω:

$$V_t^{\hat{z},\mathrm{in}}(\mathbf{a}_{1:t-1}, \omega; T, \mathbf{y}) = \max_{a \in \mathcal{A}} \left\{ r_t(\mathbf{a}_{1:t-1} \oplus a, \omega) - \hat{z}_t(\mathbf{a}_{1:t-1} \oplus a, \omega; T, \mathbf{y}) + V_{t+1}^{\hat{z},\mathrm{in}}(\mathbf{a}_{1:t-1} \oplus a, \omega; T, \mathbf{y}) \right\}, \tag{A.102}$$

with $V_{T+1}^{\hat{z},\mathrm{in}}(\cdot, \omega; T, \mathbf{y}) = 0$. Then, we have $W^{\hat{z}}(T, \mathbf{y}) = \mathbb{E}\left[V_1^{\hat{z},\mathrm{in}}(\emptyset, \omega; T, \mathbf{y}) \right]$. To prove the claim, it suffices to show that, for any given ω,

$$V_t^{\hat{z},\mathrm{in}}\left(\mathbf{a}_{1:t-1}, \omega; T, \mathbf{y}_{t-1}(\mathbf{a}_{1:t-1}, \omega; \mathbf{y})\right) \leq \widehat{V}\left(T - t + 1, \mathbf{y}_{t-1}(\mathbf{a}_{1:t-1}, \omega; \mathbf{y})\right), \tag{A.103}$$

for all $\mathbf{a}_{1:t-1} \in \mathcal{A}^{t-1}$ and for all $t = 1, \cdots, T+1$.

We argue by induction. As a terminal case, when $t = T + 1$, the inequality (A.103) holds trivially since both sides are zero. Fix t and suppose that the inequality (A.103) holds for $t + 1$. Omitting ω for brevity, we get

$$\widehat{V}\left(T - t + 1, \mathbf{y}_{t-1}(\mathbf{a}_{1:t-1})\right) - V_t^{\hat{z},\mathrm{in}}\left(\mathbf{a}_{1:t-1}; T, \mathbf{y}_{t-1}(\mathbf{a}_{1:t-1})\right) \tag{A.104}$$

$$= \widehat{V}\left(T - t + 1, \mathbf{y}_{t-1}(\mathbf{a}_{1:t-1})\right) - \max_{a \in \mathcal{A}} \left\{ r_t(\mathbf{a}_{1:t-1} \oplus a) - \hat{z}_t(\mathbf{a}_{1:t-1} \oplus a; T, \mathbf{y}) + V_{t+1}^{\hat{z},\mathrm{in}}(\mathbf{a}_{1:t-1} \oplus a; T, \mathbf{y}) \right\} \tag{A.105}$$

$$= \min_{a \in \mathcal{A}} \left\{ \begin{array}{l} \underbrace{\widehat{V}\left(T - t, \mathbf{y}_t(\mathbf{a}_{1:t})\right) - V_{t+1}^{\hat{z},\mathrm{in}}(\mathbf{a}_{1:t-1} \oplus a; T, \mathbf{y})}_{\geq 0 \quad (\because \text{ induction hypothesis})} \\ \underbrace{+ \widehat{V}\left(T - t + 1, \mathbf{y}_{t-1}(\mathbf{a}_{1:t-1})\right) - \mathbb{E}\left[r_t(\mathbf{a}_{1:t-1} \oplus a) + \widehat{V}\left(T - t, \mathbf{y}_t(\mathbf{a}_{1:t-1} \oplus a)\right) \middle| H_{t-1} \right]}_{\geq 0 \quad (\because \text{ Remark A.4.1})} \end{array} \right\} \tag{A.106}$$

$$\geq 0. \tag{A.107}$$

□

Proof of "$W^{\mathrm{TS}}(T, \mathbf{y}) \geq W^{\text{IRS.V-EMAX}}(T, \mathbf{y})$." Recall that $z_t^{\text{IRS.V-EMAX}}$ is a penalty function generated by W^{TS}. We observe that $W^{\mathrm{TS}}(\cdot, \cdot)$ is a supersolution: for any T and $\mathbf{y}$,

$$W^{\mathrm{TS}}(T, \mathbf{y}) = \mathbb{E}_{\mathbf{y}}\left[T \times \max_{a \in \mathcal{A}} \mu_a(\theta_a)\right] \tag{A.108}$$

$$= \mathbb{E}_{\mathbf{y}}\left[\max_{a \in \mathcal{A}} \mu_a(\theta_a)\right] + W^{\mathrm{TS}}(T-1, \mathbf{y}) \tag{A.109}$$

$$\geq \max_{a \in \mathcal{A}} \left\{\mathbb{E}_{y_a}\left[\mu_a(\theta_a)\right] + W^{\mathrm{TS}}(T-1, \mathbf{y})\right\} \tag{A.110}$$

$$= \max_{a \in \mathcal{A}} \left\{\mathbb{E}_{\mathbf{y}}\left[R_{a,1} + W^{\mathrm{TS}}(T-1, \mathbf{y})\right]\right\} \tag{A.111}$$

$$= \max_{a \in \mathcal{A}} \left\{\mathbb{E}_{\mathbf{y}}\left[R_{a,1} + W^{\mathrm{TS}}(T-1, \mathcal{U}(\mathbf{y}, a, R_{a,1}))\right]\right\}. \tag{A.112}$$

The last equality holds since $\mathbb{E}\left[W^{\mathrm{TS}}\left(T-1, \mathcal{U}(\mathbf{y}, a_1, r_1(a_1, \omega))\right)\right] = W^{\mathrm{TS}}(T-1, \mathbf{y})$, as argued in (2.39). By Lemma A.4.2, we have $W^{\text{IRS.V-EMAX}}(T, \mathbf{y}) \leq W^{\mathrm{TS}}(T, \mathbf{y})$ which also holds in a stronger sense. □

A.4.4 Proof of Theorem 2.4.2

Suboptimality decomposition

As in §A.3.1, we define the Q-values of the inner problem given a particular outcome ω, a penalty function $z_t(\cdot)$, a time horizon T, and a prior belief $\mathbf{y}$.

$$Q_t^{z,\text{in}}(\mathbf{a}_{1:t-1}, a, \omega; T, \mathbf{y}) = r_t(\mathbf{a}_{1:t-1} \oplus a, \omega) - z_t(\mathbf{a}_{1:t-1} \oplus a, \omega; T, \mathbf{y}) \tag{A.113}$$

$$+ V_{t+1}^{z,\text{in}}(\mathbf{a}_{1:t-1} \oplus a, \omega; T, \mathbf{y}),$$

$$V_t^{z,\text{in}}(\mathbf{a}_{1:t-1}, \omega; T, \mathbf{y}) = \max_{a \in \mathcal{A}} \left\{Q_t^{z,\text{in}}(\mathbf{a}_{1:t-1}, a, \omega; T, \mathbf{y})\right\}, \tag{A.114}$$

with $V_{T+1}^{z,\text{in}}(\cdot, \omega; T, \mathbf{y}) \equiv 0$. Additionally define the total payoff of an action sequence and the hindsight best action under penalties:

$$\mathcal{S}^z(\mathbf{a}_{1:T}, \omega; T, \mathbf{y}) \triangleq \sum_{t=1}^{T} r_t(\mathbf{a}_{1:t}, \omega) - z_t(\mathbf{a}_{1:t}, \omega; T, \mathbf{y}), \tag{A.115}$$

$$a_t^{z,*}(\mathbf{a}_{1:t-1}, \omega; T, \mathbf{y}) \triangleq \operatorname*{argmax}_{a \in \mathcal{A}} \left\{ Q_t^{z,\text{in}}(\mathbf{a}_{1:t-1}, a, \omega; T, \mathbf{y}) \right\}. \tag{A.116}$$

We have $V_1^{z,\text{in}}(\emptyset, \omega; T, \mathbf{y}) = \max_{\mathbf{a}_{1:T} \in \mathcal{A}^T} \mathcal{S}^z(\mathbf{a}_{1:T}, \omega; T, \mathbf{y})$.

Proposition A.4.3 (Suboptimality decomposition)**.** *Given a non-anticipating policy $\pi \in \Pi_{\mathbb{F}}$ and a dual-feasible penalty function z_t, the suboptimality gap is the sum of the instantaneous suboptimalities of individual actions taken by π along the sample path: i.e.,*

$$W^z(T, \mathbf{y}) - V(\pi, T, \mathbf{y}) = \mathbb{E}_{\mathbf{y}} \left[\max_{\mathbf{a}_{1:T}} \{\mathcal{S}^z(\mathbf{a}_{1:T}, \omega; T, \mathbf{y})\} - \mathcal{S}^z(\mathbf{A}_{1:T}^{\pi}, \omega; T, \mathbf{y}) \right] \tag{A.117}$$

$$= \mathbb{E}_{\mathbf{y}} \left[\sum_{t=1}^{T} \max_{a} \left\{ Q_t^{z,\text{in}}(\mathbf{A}_{1:t-1}^{\pi}, a, \omega; T, \mathbf{y}) \right\} - Q_t^{z,\text{in}}(\mathbf{A}_{1:t-1}^{\pi}, A_t^{\pi}, \omega; T, \mathbf{y}) \right], \tag{A.118}$$

where the expectation is taken with respect to the randomness of outcome ω and the randomness of policy π.

Proof. The first equality immediately follows from the definition of W^z and mean equivalence (Proposition A.3.1). Now fix ω, T, and $\mathbf{y}$. Consider the (pathwise) suboptimality of the action sequence $\mathbf{A}_{1:T}^{\pi}$ compared to the clairvoyant optimal solution. It can be decomposed into the instantaneous suboptimalty incurred by the individual action at each time:

$$\max_{\mathbf{a}_{1:T}} \{\mathcal{S}^z(\mathbf{a}_{1:T})\} - \mathcal{S}^z(\mathbf{A}_{1:T}^{\pi}) = \sum_{t=1}^{T} \max_{a} \left\{ Q_t^{z,\text{in}}(\mathbf{A}_{1:t-1}^{\pi}, a) \right\} - Q_t^{z,\text{in}}(\mathbf{A}_{1:t-1}^{\pi}, A_t^{\pi}). \tag{A.119}$$

By taking expectation, we obtain the second equality. □

The next lemma shows that the instantaneous suboptimalty of the first action can be expressed

in terms of mean reward metrics for each of the IRS penalty functions.

Lemma A.4.3. *Fix time horizon T, prior belief* $\mathbf{y}$*, and the true outcome ω, and hide the dependency on them in notation for $Q_1^{z,\mathrm{in}}(\cdot)$, $a_1^{z,*}(\cdot)$, $\mu_a(\cdot)$ and $\hat{\mu}_{a,n}(\cdot)$. For each of the penalty functions z^{TS}, $z^{\textsc{Irs.FH}}$, and $z^{\textsc{Irs.V-Zero}}$, the instantaneous suboptimalty of action $a \in \mathcal{A}$ satisfies the following:*

(1) When $z \equiv z^{\mathrm{TS}}$,

$$Q_1^{z,\mathrm{in}}(a_1^{z,*}) - Q_1^{z,\mathrm{in}}(a) = \mu_{a_1^{z,*}} - \mu_a. \tag{A.120}$$

(2) When $z \equiv z^{\textsc{Irs.FH}}$,

$$Q_1^{z,\mathrm{in}}(a_1^{z,*}) - Q_1^{z,\mathrm{in}}(a) = \hat{\mu}_{a_1^{z,*},T-1} - \hat{\mu}_{a,T-1}. \tag{A.121}$$

(3) When $z \equiv z^{\text{V-Zero}}$,

$$Q_1^{z,\mathrm{in}}(a_1^{z,*}) - Q_1^{z,\mathrm{in}}(a) \leq \max_{0 \leq n \leq T-1} \left\{ \hat{\mu}_{a_1^{z,*},n} \right\} - \hat{\mu}_{a,0}. \tag{A.122}$$

Proof. (1) When $z \equiv z^{\mathrm{TS}}$, we have

$$Q_1^{z,\mathrm{in}}(a) = \mu_a + (T-1) \times \max_{a'} \mu_{a'}. \tag{A.123}$$

Since the last term does not depend on action a, the claim follows.

(2) When $z \equiv z^{\textsc{Irs.FH}}$, we obtain the claim by replacing μ_a with $\hat{\mu}_{a,T-1}$ in the above proof.

(3) When $z \equiv z^{\textsc{Irs.V-Zero}}$, recall that the associated inner problem is to find an optimal allocation: i.e.,

$$\max_{\mathbf{n}_{1:K} \in N_T} \left\{ \sum_{a=1}^{K} \sum_{i=0}^{n_a - 1} \hat{\mu}_{a,i} \right\}. \tag{A.124}$$

Let $\mathbf{n}^*_{1:K}$ be the optimal allocation. Observe that the suboptimality is incurred only when $n^*_a = 0$, it is no worse than $\hat{\mu}_{a^*,n^*_{a^*}} - \hat{\mu}_{a,0}$ (the loss if the payoff when pulling a one more time but pulling $a_1^{z,*}$ one less time). Since $n^*_{a^*} \leq T-1$, the claim follows. □

Recursive structure of IRS penalty functions

To describe the recursive structure of Bayesian MAB problems explicitly, we define a shift operator $\mathcal{M}_t : \mathcal{A}^t \times \Omega \mapsto \Omega$,

$$\mathcal{M}_t(\mathbf{a}_{1:t}, \omega) \triangleq \left(R_{a,n_a}; \forall n_a > n_t(\mathbf{a}_{1:t}, a), \forall a \in \mathcal{A}\right). \tag{A.125}$$

The shifted outcome $\mathcal{M}_{t-1}(\mathbf{a}_{1:t-1}, \omega)$ encodes the remaining reward realizations after taking $\mathbf{a}_{1:t-1}$.

Remark A.4.2 (Recursive structure of remaining uncertainties)**.** *Conditioned on $\mathcal{H}_{t-1}(\mathbf{a}_{1:t-1}, \omega)$, the remaining uncertainties are sufficiently described by $\mathbf{y}_{t-1}(\mathbf{a}_{1:t-1}, \omega; \mathbf{y})$, i.e.,*

$$\mathcal{M}_{t-1}(\mathbf{a}_{1:t-1}, \omega) | \, H_{t-1}(\mathbf{a}_{1:t-1}, \omega) \quad \sim \quad \mathcal{I}\left(\mathbf{y}_{t-1}(\mathbf{a}_{1:t-1}, \omega; \mathbf{y})\right). \tag{A.126}$$

Remark A.4.3 (Recursive structure of IRS penalties)**.** *Each of penalty functions* (2.22)–(2.26) *has the following form:*

$$z_t(\mathbf{a}_{1:t}, \omega; T, \mathbf{y}) = \varphi^z(\mathcal{M}_{t-1}(\mathbf{a}_{1:t-1}, \omega), T - t + 1, \mathbf{y}_{t-1}(\mathbf{a}_{1:t-1}, \omega; \mathbf{y})), \tag{A.127}$$

for some function $\varphi^z : \Omega \times \mathbb{N} \times \mathcal{Y} \mapsto \mathbb{R}$, i.e., the penalty at each time is completely determined by the remaining rewards $\mathcal{M}_{t-1}(\mathbf{a}_{1:t-1}, \omega)$, the remaining time horizon $T - t + 1$, and the prior belief $\mathbf{y}_{t-1}(\mathbf{a}_{1:t-1}, \omega)$ at that moment.

Remark A.4.2 immediately follows from Bayes' rule, and Remark A.4.3 can be easily verified. We observe the recursive structure of the sequential inner problems that the DM solves throughout the decision-making process, which can be characterized by the following property.

Proposition A.4.4 (Generalized posterior sampling)**.** *For each of penalty functions* (2.22)–(2.26)*, the IRS policy π is randomized in such a way that it takes an action a with the probability that the*

action a is indeed the best action $a_t^{z,*}$ at that moment, i.e.,

$$\mathbb{P}\left[A_t^\pi = a \middle| \mathcal{F}_{t-1}\right] = \mathbb{P}\left[a_t^{z,*}(\mathbf{A}_{1:t-1}^\pi, \omega) = a \middle| \mathcal{F}_{t-1}\right], \quad \forall a, \quad \forall t. \tag{A.128}$$

The source of uncertainty in the LHS is the randomness of the policy (embedded in $\tilde{\omega}$) and that in the RHS is the randomness of nature (embedded in ω). Here we assume that the tie-breaking rule in argmax *of* (A.116) *is identical to the one used when π^z solves the inner problem.*

Proof. Observe that the IRS's action A_t^π can be represented as

$$A_t^\pi = a_1^{z,*}\left(\emptyset, \tilde{\omega}; T - t + 1, \mathbf{y}_{t-1}(\mathbf{A}_{1:t-1}^\pi, \omega; \mathbf{y})\right), \tag{A.129}$$

where $\tilde{\omega} \sim \mathcal{I}(\mathbf{y}_{t-1}(\mathbf{A}_{1:t-1}^\pi, \omega; \mathbf{y}))$, i.e., the action that the clairvoyant DM will take in an MAB instance specified by horizon $T-t+1$, prior belief $\mathbf{y}_{t-1}(\mathbf{A}_{1:t-1}^\pi, \omega; \mathbf{y})$, and the outcome $\tilde{\omega}$. Therefore, it suffices to verify that the inner problem that π solves at time t is identically distributed with the sub-inner problem with respect to ground-truth ω (i.e., the subproblem given the past action sequence $\mathbf{A}_{1:t-1}^\pi$).

Fix time t, past actions $\mathbf{a}_{1:t-1} = \mathbf{A}_{1:t-1}^\pi$, and the true outcome ω. The sub-inner problem determining $a_t^{z,*}(\mathbf{a}_{1:t-1}, \omega)$ is

$$\max_{\mathbf{a}'_{t:T}} \left\{ \sum_{s=t}^{T} r_s(\mathbf{a}_{1:t-1} \oplus \mathbf{a}'_{t:s}, \omega) - z_s(\mathbf{a}_{1:t-1} \oplus \mathbf{a}'_{t:s}, \omega; T, \mathbf{y}) \right\}. \tag{A.130}$$

By Remark A.4.3, for any $s \in \{t, \ldots, T\}$, the penalty at (inner) time s is given by

$$z_s(\mathbf{a}_{1:t-1} \oplus \mathbf{a}'_{t:s}, \omega; T, \mathbf{y}) \tag{A.131}$$

$$= \varphi^z(\mathcal{M}_{s-1}(\mathbf{a}_{1:t-1} \oplus \mathbf{a}'_{t:s-1}, \omega), T - s + 1, \mathbf{y}_{s-1}(\mathbf{a}_{1:t-1} \oplus \mathbf{a}'_{t:s-1}, \omega; \mathbf{y})) \tag{A.132}$$

$$= \varphi^z \begin{pmatrix} \mathcal{M}_{s-t}(\mathbf{a}'_{t:s-1}, \mathcal{M}_{t-1}(\mathbf{a}_{1:t-1}, \omega)), \\ (T - t + 1) - (s - t), \\ \mathbf{y}_{s-t}(\mathbf{a}'_{t:s-1}, \mathcal{M}_{t-1}(\mathbf{a}_{1:t-1}, \omega); \mathbf{y}_{t-1}(\mathbf{a}_{1:t-1}, \omega; \mathbf{y}) \end{pmatrix} \tag{A.133}$$

$$= z_{s-t+1}(\mathbf{a}'_{t:s}, \mathcal{M}_{t-1}(\mathbf{a}_{1:t-1}, \omega); T - t + 1, \mathbf{y}_{t-1}(\mathbf{a}_{1:t-1}, \omega; \mathbf{y})). \tag{A.134}$$

For rewards, similarly, we have $r_s(\mathbf{a}_{1:t-1} \oplus \mathbf{a}'_{t:s}, \omega) = r_{s-t+1}(\mathbf{a}'_{t:s}, \mathcal{M}_{t-1}(\mathbf{a}_{1:t-1}, \omega))$. Therefore, the sub-inner problem (A.130) is reformulated as

$$\max_{\mathbf{a}'_{t:T}} \left\{ \sum_{s=t}^{T} r_{s-t+1}(\mathbf{a}'_{t:s}, \mathcal{M}_{t-1}(\mathbf{a}_{1:t-1}, \omega)) - z_{s-t+1}(\mathbf{a}'_{t:s}, \mathcal{M}_{t-1}(\mathbf{a}_{1:t-1}, \omega); T - t + 1, \mathbf{y}_{t-1}(\mathbf{a}_{1:t-1}, \omega; \mathbf{y})) \right\}. \tag{A.135}$$

Given the fact that the shifted outcome $\mathcal{M}_{t-1}(\mathbf{a}_{1:t-1}, \omega)$ and the sampled outcome $\tilde{\omega}$ are identically distributed with $\mathcal{I}(\mathbf{y}_{t-1}(\mathbf{a}_{1:t-1}, \omega; \mathbf{y}))$ conditionally on $H_{t-1}(\mathbf{a}_{1:t-1}, \omega)$ (Remark A.4.2), this sub-inner problem follows the same distribution with

$$\max_{\mathbf{a}'_{1:T-t+1}} \left\{ \sum_{s=1}^{T-t+1} r_s(\mathbf{a}'_{1:s}, \tilde{\omega}) - z_s(\mathbf{a}'_{1:s}, \tilde{\omega}, T - t + 1, \mathbf{y}_{t-1}(\mathbf{a}_{1:t-1}, \omega; \mathbf{y})) \right\}, \tag{A.136}$$

which characterizes the IRS's action A_t^π. Therefore, $a_t^{z,*}(\mathbf{A}_{1:t-1}^\pi, \omega)$ is identically distributed with A_t^π conditioned on $\mathcal{F}_{t-1}$. □

Remark A.4.4. *Utilizing the recursive structure of IRS penalty functions, Lemma A.4.3 can be extended to describe the instantaneous suboptimality of the t^{th} action. Fix true outcome ω and past actions $\mathbf{a}_{1:t-1}$, and hide the dependency on them in notation for $Q_t^{z,\text{in}}(\cdot)$, $a_t^{z,*}(\cdot)$, $n_t(\cdot)$, $\mu_a(\cdot)$ and $\hat{\mu}_{a,n}(\cdot)$.*

(1) When $z \equiv z^{\mathrm{TS}}$,

$$Q_t^{z,\mathrm{in}}(a_t^{z,*}) - Q_t^{z,\mathrm{in}}(a) = \mu_{a_t^{z,*}} - \mu_a. \tag{A.137}$$

(2) When $z \equiv z^{\textsc{Irs.FH}}$,

$$Q_t^{z,\mathrm{in}}(a_t^{z,*}) - Q_t^{z,\mathrm{in}}(a) = \hat{\mu}_{a_t^{z,*},n_{t-1}(a_t^{z,*})+T-t} - \hat{\mu}_{a,n_{t-1}(a)+T-t}. \tag{A.138}$$

(3) When $z \equiv z^{\text{V-}\textsc{Zero}}$,

$$Q_t^{z,\mathrm{in}}(a_t^{z,*}) - Q_t^{z,\mathrm{in}}(a) \leq \max_{0\leq n\leq T-t} \left\{\hat{\mu}_{a_1^{z,*},n_{t-1}(a_1^{z,*})+n}\right\} - \hat{\mu}_{a,n_{t-1}(a)}. \tag{A.139}$$

Preliminary lemmas on MAB with natural exponential family distributions

We first describe the notion of sub-Gaussian random variable as an effective tool for bounding its tail behavior.

Definition A.4.2 (Sub-Gaussian random variable)**.** *A random variable* X *is* σ*-sub-Gaussian if*

$$\mathbb{E}\left[\exp\left(\lambda(X - \mathbb{E}X)\right)\right] \leq \exp\left(\frac{\sigma\lambda^2}{2}\right), \quad \forall\lambda \in \mathbb{R}, \tag{A.140}$$

for some $\sigma > 0$.

Lemma A.4.4. *Given a random variable* X*, suppose that there exists* $\sigma > 0$ *such that*

$$\mathbb{P}\left[X \geq \mathbb{E}X + z\sigma\right] \leq e^{-z^2/2}, \quad \forall z \geq 0. \tag{A.141}$$

Then, the following holds:

$$\mathbb{E}\left[\left(X - (\mathbb{E}X + z\sigma)\right)^+\right] \leq \frac{\sigma}{z}e^{-z^2/2}, \quad \forall z > 0. \tag{A.142}$$

Corollary A.4.3. *If a random variable* X *is* σ*-sub-Gaussian, it satisfies the condition of Lemma A.4.4 and hence the inequality* (A.142) *holds.*

Proof. With $\mu \triangleq \mathbb{E}X$, we have

$$\mathbb{E}\left[(X - (\mu + z\sigma))^+\right] = \int_{x=\mu+z\sigma}^{\infty} \mathbb{P}\left[X \geq x\right] dx = \int_{t=z}^{\infty} \mathbb{P}\left[X \geq \mu + t\sigma\right] \sigma dt \leq \sigma \int_{t=z}^{\infty} e^{-t^2/2} dt. \tag{A.143}$$

Utilizing the tail bound established for the standard normal distribution, we can show that

$$\int_{t=z}^{\infty} \frac{1}{\sqrt{2\pi}} e^{-t^2/2} dt \leq \frac{1}{z} \frac{e^{-z^2/2}}{\sqrt{2\pi}}. \tag{A.144}$$

By combining these two inequalities, we obtain the desired result.

The corollary simply follows from Markov inequality: for any $z \geq 0$ and $\lambda \geq 0$, we have

$$\mathbb{P}[X \geq \mu + z\sigma] = \mathbb{P}\left[e^{\lambda(X-\mu)} \geq e^{\lambda z\sigma}\right] \leq \frac{\mathbb{E}[e^{\lambda(X-\mu)}]}{e^{\lambda x\sigma}} \leq \exp\left(\frac{\sigma^2\lambda^2}{2} - \lambda z\sigma\right). \tag{A.145}$$

By taking $\lambda = \frac{z}{\sigma}$, it follows that $\mathbb{P}[X \geq \mu + z\sigma] \leq e^{-z^2/2}$. □

We now return to the context of MAB problems and show that the mean reward metrics are sub-Gaussian.

Lemma A.4.5 (Sub-Gaussianity of mean reward metrics)**.** *Consider the setting of Theorem 2.4.2, i.e., the reward distribution of arm a is described by an L-smooth log-partition function $A_a(\theta_a)$ and hyper-parameters (ξ_a, ν). Then, the conditional mean reward μ_a is $\sqrt{L/\nu}$-sub-Gaussian: i.e.,*

$$\mathbb{E}_{(\xi_a,\nu)}\left[\exp\left(\lambda(\mu_a - \bar{\mu}_a)\right)\right] \leq \exp\left(\frac{L\lambda^2}{2\nu}\right), \quad \forall \lambda \in \mathbb{R}, \tag{A.146}$$

where $\bar{\mu}_a = \mathbb{E}_{(\xi_a,\nu)}[\mu_a] = \frac{\xi_a}{\nu}$ is the prior predictive mean reward (i.e., the unconditional mean reward). Furthermore, the posterior predictive mean reward $\hat{\mu}_{a,n}$ is $\sqrt{\frac{Ln}{\nu(\nu+n)}}$-sub-Gaussian: i.e.,

$$\mathbb{E}_{(\xi_a,\nu)}\left[\exp\left(\lambda(\hat{\mu}_{a,n} - \bar{\mu}_a)\right)\right] \leq \exp\left(\frac{\lambda^2}{2} \times \frac{Ln}{\nu(\nu+n)}\right), \quad \forall \lambda \in \mathbb{R}. \tag{A.147}$$

Proof. We first prove that μ_a is $\sqrt{L/\nu}$-sub-Gaussian. Due to L-smoothness condition, $A_a(\theta_a)$ is

finite valued for all $\theta_a \in \mathbb{R}$. For any $\lambda \in \mathbb{R}$, we have

$$\mathbb{E}_{(\xi_a,\nu)}\left[\exp\left(\lambda\mu_a\right)\right] \overset{(i)}{=} \mathbb{E}_{(\xi_a,\nu)}\left[\exp\left(\lambda A_a'(\theta_a)\right)\right] \tag{A.148}$$

$$= \int_{-\infty}^{\infty} \exp\left(\lambda A_a'(\theta_a)\right) \times f_a(\xi_a,\nu)\exp\left(\xi_a\theta_a - \nu A_a(\theta_a)\right) d\theta_a \tag{A.149}$$

$$= \int_{-\infty}^{\infty} f_a(\xi_a,\nu)\exp\left\{\xi_a\theta_a - \nu A_a(\theta_a) + \lambda A_a'(\theta_a)\right\} d\theta_a \tag{A.150}$$

$$= \int_{-\infty}^{\infty} f_a(\xi_a,\nu)\exp\left\{\xi_a\theta_a - \nu\left(A_a(\theta_a) - \lambda/\nu\cdot A_a'(\theta_a)\right)\right\} d\theta_a \tag{A.151}$$

$$\overset{(ii)}{\leq} \int_{-\infty}^{\infty} f_a(\xi_a,\nu)\exp\left\{\xi_a\theta_a - \nu\left(A_a(\theta_a - \lambda/\nu) - \frac{L\lambda^2}{2\nu^2}\right)\right\} d\theta_a \tag{A.152}$$

$$= \exp\left(\frac{L\lambda^2}{2\nu}\right) \times \int_{-\infty}^{\infty} f_a(\xi_a,\nu)\exp\left\{\xi_a\theta_a - \nu A_a(\theta_a - \lambda/\nu)\right\} d\theta_a \tag{A.153}$$

$$= \exp\left(\frac{\xi_a\lambda}{\nu} + \frac{L\lambda^2}{2\nu}\right) \times \int_{-\infty}^{\infty} f_a(\xi_a,\nu)\exp\left\{\xi_a(\theta_a - \lambda/\nu) - \nu A_a(\theta_a - \lambda/\nu)\right\} d\theta_a \tag{A.154}$$

$$= \exp\left(\frac{\xi_a\lambda}{\nu} + \frac{L\lambda^2}{2\nu}\right) \times \int_{-\infty}^{\infty} f_a(\xi_a,\nu)\exp\left\{\xi_a\theta_a - \nu A_a(\theta_a)\right\} d\theta_a \tag{A.155}$$

$$= \exp\left(\frac{\xi_a\lambda}{\nu} + \frac{L\lambda^2}{2\nu}\right), \tag{A.156}$$

where we have utilized that (i) $\mu_a(\theta_a) = A_a'(\theta_a)$ and (ii) $A_a(\theta_a + \delta) \leq A_a(\theta_a) + \delta A_a'(\theta_a) + \frac{L}{2}\delta^2$. Since $\bar{\mu}_a = \xi_a/\nu$, we obtained the desired result.

Next we focus on the posterior predictive mean reward $\hat{\mu}_{a,n}$. Recall that we have

$$\hat{\mu}_{a,n} = \frac{\xi_a + \sum_{i=1}^{n} R_{a,i}}{\nu + n}. \tag{A.157}$$

For any $\lambda \in \mathbb{R}$, we have

$$\mathbb{E}_{(\xi_a,\nu)}\left[\exp\left(\lambda \sum_{i=1}^{n} R_{a,i}\right)\right] = \mathbb{E}_{(\xi_a,\nu)}\left[\mathbb{E}\left\{\exp\left(\lambda \sum_{i=1}^{n} R_{a,i}\right)\middle| \theta_a\right\}\right] \tag{A.158}$$

$$\overset{(i)}{=} \mathbb{E}_{(\xi_a,\nu)}\left[\mathbb{E}\left\{\exp\left(\lambda R_{a,1}\right)\middle| \theta_a\right\}^n\right] \tag{A.159}$$

$$\overset{(ii)}{=} \mathbb{E}_{(\xi_a,\nu)}\left[\exp\left\{A_a\left(\theta_a+\lambda\right) - A_a(\theta_a)\right\}^n\right] \tag{A.160}$$

$$\overset{(iii)}{\leq} \mathbb{E}_{(\xi_a,\nu)}\left[\exp\left\{\lambda \cdot A_a'(\theta_a) + \frac{L\lambda^2}{2}\right\}^n\right] \tag{A.161}$$

$$\overset{(iv)}{=} \mathbb{E}_{(\xi_a,\nu)}\left[\exp\left\{n\lambda \cdot \mu_a + \frac{Ln\lambda^2}{2}\right\}\right] \tag{A.162}$$

$$= \exp\left(n\lambda\bar{\mu}_a + \frac{Ln\lambda^2}{2}\right) \times \mathbb{E}_{(\xi_a,\nu)}\left[\exp\left(n\lambda(\mu_a - \bar{\mu}_a)\right)\right] \tag{A.163}$$

$$\overset{(v)}{\leq} \exp\left(n\lambda\bar{\mu}_a + \frac{Ln\lambda^2}{2}\right) \times \exp\left(\frac{Ln^2\lambda^2}{2\nu}\right) \tag{A.164}$$

$$= \exp\left(n\lambda\bar{\mu}_a\right) \times \exp\left(\frac{\lambda^2}{2} \times \frac{Ln(\nu+n)}{\nu}\right), \tag{A.165}$$

where we have utilized that (i) $R_{a,i}$'s are conditionally independent given θ_a, (ii) the moment-generating function of $R_{a,1}$ is given by $\mathbb{E}[\lambda R_a|\theta_a] = \exp\left(A_a(\theta_a+\lambda) - A_a(\theta_a)\right)$, (iii) $A_a(\cdot)$ is L-smooth, (iv) $A_a'(\theta_a) = \mu_a(\theta_a)$, and (v) μ_a is $\sqrt{L/\nu}$-sub-Gaussian. Given that $\mathbb{E}\left[\sum_{i=1}^{n} R_{a,i}\right] = n\bar{\mu}_a$, we just have shown that the sum $\sum_{i=1}^{n} R_{a,i}$ is $\sqrt{\frac{Ln(\nu+n)}{\nu}}$-sub-Gaussian. Therefore, its scaled version $\frac{\sum_{i=1}^{n} R_{a,i}}{\nu+n}$ is $\sqrt{\frac{Ln}{\nu(\nu+n)}}$-sub-Gaussian, and so is $\hat{\mu}_{a,n}$. □

Lemma A.4.6. *Consider the setting of Theorem 2.4.2. With $\sigma_n \triangleq \sqrt{\frac{Ln}{\nu(\nu+n)}}$, the following holds:*

$$\mathbb{E}\left[\left(\max_{0\leq i\leq n} \hat{\mu}_{a,i} - (\bar{\mu}_a + z\sigma_n)\right)^+\right] \leq \frac{\sigma_n}{z} e^{-z^2/2}, \quad \forall z > 0. \tag{A.166}$$

Proof. Recall that the posterior predictive mean reward process $\{\hat{\mu}_{a,n}\}_{n\geq 0}$ is the martingale with respect to the filtration generated by reward realizations $R_{a,1}, R_{a,2}, \ldots$ and whose mean is $\bar{\mu}_a$. Therefore, $\{\exp(\lambda\hat{\mu}_{a,n})\}_{n\geq 0}$ is a positive submartingale for any given $\lambda \geq 0$. By Doob's maximal

inequality, we deduce that

$$\mathbb{P}\left[\max_{0\le i\le n}\hat{\mu}_{a,i}\ge\bar{\mu}_a+z\sigma_n\right]=\mathbb{P}\left[\max_{0\le i\le n}\exp\left(\lambda(\hat{\mu}_{a,i}-\bar{\mu}_a)\right)\ge\exp\left(\lambda z\sigma_n\right)\right]\le\frac{\mathbb{E}\left[\exp\left(\lambda(\hat{\mu}_{a,n}-\bar{\mu}_a)\right)\right]}{\exp\left(\lambda z\sigma_n\right)}. \tag{A.167}$$

By Lemma A.4.5, since $\hat{\mu}_{a,n}$ is σ_n-sub-Gaussian, we further have

$$\frac{\mathbb{E}\left[\exp\left(\lambda(\hat{\mu}_{a,n}-\bar{\mu}_a)\right)\right]}{\exp\left(\lambda z\sigma_n\right)}\le\frac{\exp\left(\frac{\lambda^2\sigma_n^2}{2}\right)}{\exp\left(\lambda z\sigma_n\right)}=\exp\left(\frac{\lambda^2\sigma_n^2}{2}-\lambda z\sigma_n\right). \tag{A.168}$$

Therefore, by taking $\lambda\triangleq\frac{z}{\sigma_n}$, we have $\mathbb{P}\left[\max_{0\le i\le n}\hat{\mu}_{a,i}\ge\bar{\mu}_a+z\sigma_n\right]\le e^{-z^2/2}$, and by invoking Lemma A.4.4, we obtain the claim. □

Proof of Theorem 2.4.2

Lemma A.4.7. *Consider one of the IRS penalty functions z^{TS}, $z^{\mathrm{IRS.FH}}$, and $z^{\mathrm{IRS.V\text{-}ZERO}}$. As discussed in Remark A.4.4, we have*

$$Q_t^{z,\mathrm{in}}(\mathbf{a}_{1:t-1},a_t^{z,*},\omega)-Q_t^{z,\mathrm{in}}(\mathbf{a}_{1:t-1},a,\omega)\le\mu_t^U(\mathbf{a}_{1:t-1},a_t^{z,*},\omega)-\mu_t^L(\mathbf{a}_{1:t-1},a,\omega), \tag{A.169}$$

for some $\mu_t^U(\mathbf{a}_{1:t-1},a_1^{z,*},\omega)$ and $\mu_t^L(\mathbf{a}_{1:t-1},a,\omega)$, where $a_t^{z,*}$ abbreviates $a_t^{z,*}(\mathbf{a}_{1:t-1},\omega)$. Suppose that there exists a sequence of confidence intervals $\{(L_t(a),U_t(a))\}_{a\in\mathcal{A},t\in\mathbb{N}}$ such that $(L_t(\cdot),U_t(\cdot))$ is $\sigma(H_{t-1})$-measurable, and

$$\mathbb{E}_{\mathbf{y}}\left[\left(\mu_t^U(\mathbf{a}_{1:t-1},a,\omega)-U_t(a)\right)^+\middle|H_{t-1}(\mathbf{a}_{1:t-1},\omega)\right]\le\frac{C_U}{T},\quad\forall a,\forall t \tag{A.170}$$

$$\mathbb{E}_{\mathbf{y}}\left[\left(L_t(a)-\mu_t^L(\mathbf{a}_{1:t-1},a,\omega)\right)^+\middle|H_{t-1}(\mathbf{a}_{1:t-1},\omega)\right]\le\frac{C_L}{T},\quad\forall a,\forall t \tag{A.171}$$

for some constants $C_U > 0$ and $C_L > 0$. Then, for IRS policy π induced by the chosen penalty function, we have

$$W^z(T, \mathbf{y}) - V(\pi, T, \mathbf{y}) \leq C_U + C_L + \sum_{t=1}^{T} \mathbb{E}\left[U_t(A_t^\pi) - L_t(A_t^\pi)\right]. \tag{A.172}$$

Proof. Let $A_t^* \triangleq a_t^{z,*}(\mathbf{A}_{1:t-1}^\pi, \omega)$, and let $\mathbb{E}_t[\cdot]$ denote $\mathbb{E}[\cdot|\mathcal{F}_{t-1}]$. By Proposition A.4.4 we have

$$\mathbb{E}_t[U_t(A_t^\pi)] = \sum_{a\in\mathcal{A}} U_t(a) \cdot \mathbb{P}_t[A_t^\pi = a] = \sum_{a\in\mathcal{A}} L_t(a) \cdot \mathbb{P}_t[A_t^* = a] = \mathbb{E}_t[U_t(A_t^*)]. \tag{A.173}$$

Therefore, we have

$$\begin{aligned}
&\mathbb{E}_t\left[\mu_t^U(A_t^*) - \mu_t^L(A_t^\pi)\right] && \text{(A.174)}\\
&= \mathbb{E}_t\left[\mu_t^U(A_t^*) - \mu_t^L(A_t^\pi)\right] + \mathbb{E}_t\left[U_t(A_t^\pi) - U_t(A_t^*)\right] + \mathbb{E}_t\left[L_t(A_t^\pi) - L_t(A_t^\pi)\right] && \text{(A.175)}\\
&= \mathbb{E}_t\left[\mu_t^U(A_t^*) - U_t(A_t^*)\right] + \mathbb{E}_t\left[L_t(A_t^\pi) - \mu_t^L(A_t^\pi)\right] + \mathbb{E}_t\left[U_t(A_t^\pi) - L_t(A_t^\pi)\right] && \text{(A.176)}\\
&\leq \mathbb{E}_t\left[\left(\mu_t^U(A_t^*) - U_t(A_t^*)\right)^+\right] + \mathbb{E}_t\left[\left(L_t(A_t^\pi) - \mu_t^L(A_t^\pi)\right)^+\right] + \mathbb{E}_t\left[U_t(A_t^\pi) - L_t(A_t^\pi)\right]. && \text{(A.177)}
\end{aligned}$$

We further observe that

$$\mathbb{E}_t\left[\left(\mu_t^U(A_t^*) - U_t(A_t^*)\right)^+\right] = \sum_{a\in\mathcal{A}} \mathbb{E}_t\left[\left(\mu_t^U(a) - U_t(a)\right)^+\right] \mathbb{P}_t[A_t^* = a] \leq \frac{C_U}{T} \sum_{a\in\mathcal{A}} \mathbb{P}_t[A_t^* = a] = \frac{C_U}{T}. \tag{A.178}$$

Similarly, we have $\mathbb{E}_t\left[\left(L_t(A_t^\pi) - \mu_t^L(A_t^\pi)\right)^+\right] \leq \frac{C_L}{T}$. Combining all these results, we have

$$W(T,\mathbf{y}) - V(\pi,T,\mathbf{y}) \overset{\text{Prop A.4.3}}{=} \mathbb{E}\left[\sum_{t=1}^{T} Q_t^{z,\text{in}}(A_t^*) - Q_t^{z,\text{in}}(A_t^\pi)\right] \tag{A.179}$$

$$\leq \mathbb{E}\left[\sum_{t=1}^{T} \mu_t^U(A_t^*) - \mu_t^L(A_t^\pi)\right] \tag{A.180}$$

$$= \mathbb{E}\left[\sum_{t=1}^{T} \mathbb{E}_t\left[\mu_t^U(A_t^*) - \mu_t^L(A_t^\pi)\right]\right] \tag{A.181}$$

$$\leq \mathbb{E}\left[\sum_{t=1}^{T} \left(\frac{C_U}{T} + \frac{C_L}{T} + \mathbb{E}_t\left[U_t(A_t^\pi) - L_t(A_t^\pi)\right]\right)\right] \tag{A.182}$$

$$\leq C_U + C_L + \sum_{t=1}^{T} \mathbb{E}\left[U_t(A_t^\pi) - L_t(A_t^\pi)\right]. \tag{A.183}$$

□

We are now ready to prove Theorem 2.4.2. To facilitate simpler notation, we define

$$N_{t-1}^\pi(a) \triangleq n_{t-1}(\mathbf{A}_{1:t-1}^\pi, a), \quad \hat{\mu}_t^\pi(a,n) \triangleq \hat{\mu}_{a,N_{t-1}^\pi(a)+n}, \tag{A.184}$$

which represent, respectively, the number of pulls on arm a prior to time t under policy π, and the posterior predictive mean reward process given the past actions $\mathbf{A}_{1:t-1}^\pi$. Observe that for each $a \in \mathcal{A}$, the process $\{\hat{\mu}_t^\pi(a,n)\}_{n\geq 0}$ is a martingale, as discussed Remark 2.2.1.

Further define

$$\Delta_t^\pi(a,n) \triangleq \sqrt{\frac{L}{\nu + N_{t-1}^\pi(a)} \times \frac{n}{\nu + N_{t-1}^\pi(a) + n}}, \tag{A.185}$$

which is measurable with respect to $\mathcal{F}_{t-1}$. In the context of Theorem 2.4.2, the prior/posterior of arm a at time t is described by the hyperparameters $\left(\xi_a + \sum_{i=1}^{N_{t-1}^\pi(a)} R_{a,i}, \nu + N_{t-1}^\pi(a)\right)$ that converges to μ_a, and therefore Lemma A.4.5 implies that $\hat{\mu}_t^\pi(a,n)$ is $\Delta_t(a,n)$-sub-Gaussian *conditioned* on $\mathcal{F}_{t-1}$.

(1) Suboptimality analysis for TS (2.60)**.** As discussed in Remark A.4.4, for TS, we have

$$Q_t^{z,\mathrm{in}}(a_t^{z,*}) - Q_t^{z,\mathrm{in}}(a) = \mu_{a_t^{z,*}} - \mu_a = \hat{\mu}_t^\pi(a_t^{z,*}, \infty) - \hat{\mu}_t^\pi(a, \infty). \tag{A.186}$$

We construct the confidence intervals as follows:

$$U_t(a) \triangleq \hat{\mu}_t^\pi(a, 0) + \sqrt{2\log T} \times \Delta_t^\pi(a, \infty), \quad L_t(a) \triangleq \hat{\mu}_t^\pi(a, 0) + \sqrt{2\log T} \times \Delta_t^\pi(a, \infty), \tag{A.187}$$

where $\Delta_t^\pi(a, \infty) = \lim_{n\to\infty} \Delta_t^\pi(a, n) = \sqrt{\frac{L}{\nu + N_{t-1}^\pi(a)}}$ so that μ_a is $\Delta_t^\pi(a, \infty)$-sub-Gaussian conditioned on $\mathcal{F}_{t-1}$. By Lemma A.4.4, we have

$$\mathbb{E}\left[(\mu_a - U_t(a))^+ \middle| \mathcal{F}_{t-1}\right] \le \frac{\Delta_t^\pi(a, \infty)}{\sqrt{2\log T}} e^{-\frac{2\log T}{2}} \le \frac{\sqrt{L/\nu}}{T}, \tag{A.188}$$

where we use the fact that $2\log T \ge 1$ for any $T \ge 2$. Symmetrically, we have $\mathbb{E}\left[(L_t(a) - \mu_a)^+ \middle| \mathcal{F}_{t-1}\right] \le \frac{\sqrt{L/\nu}}{T}$. By Lemma A.4.7, we have

$$W^{\mathrm{TS}}(T, \mathbf{y}) - V(\pi^{\mathrm{TS}}, T, \mathbf{y}) \le 2\sqrt{L/\nu} + \sum_{t=1}^{T} \mathbb{E}\left[U_t(A_t^\pi) - L_t(A_t^\pi)\right] \tag{A.189}$$

$$= 2\sqrt{L/\nu} + 2\sqrt{2\log T} \sum_{t=1}^{T} \Delta_t^\pi(A_t^\pi, \infty). \tag{A.190}$$

Further observe that

$$\sum_{t=1}^{T} \Delta_t^\pi(A_t^\pi, \infty) = \sum_{t=1}^{T} \sqrt{\frac{L}{\nu + N_{t-1}^\pi(A_t^\pi)}} = \sum_{a\in\mathcal{A}} \sum_{n=0}^{N_T^\pi(a)-1} \frac{\sqrt{L}}{\sqrt{\nu+n}} = \sum_{a\in\mathcal{A}} \left(\frac{\sqrt{L}}{\sqrt{\nu}} + \sum_{n=1}^{N_T^\pi(a)-1} \frac{\sqrt{L}}{\sqrt{\nu+n}} \right) \tag{A.191}$$

$$\le \sum_{a\in\mathcal{A}} \left(\frac{\sqrt{L}}{\sqrt{\nu}} + \sum_{n=1}^{N_T^\pi(a)-1} \frac{\sqrt{L}}{\sqrt{n}} \right) \le \sum_{a\in\mathcal{A}} \left(\frac{\sqrt{L}}{\sqrt{\nu}} + \int_{x=0}^{N_T^\pi(a)} \frac{\sqrt{L}}{\sqrt{x}} dx \right) = \frac{K\sqrt{L}}{\sqrt{\nu}} + 2\sqrt{L} \sum_{a\in\mathcal{A}} \sqrt{N_T^\pi(a)}. \tag{A.192}$$

By utilizing Cauchy–Schwartz inequality, we deduce that

$$\sum_{a\in\mathcal{A}}\sqrt{N_T^\pi(a)} \le \sqrt{K\sum_{a\in\mathcal{A}}N_T(a)} = \sqrt{KT}. \tag{A.193}$$

Combining all these results, we conclude that

$$W^{\mathrm{TS}}(T,\mathbf{y}) - V(\pi^{\mathrm{TS}},T,\mathbf{y}) \le 2\sqrt{L}\left[\frac{1}{\sqrt{\nu}} + \sqrt{2\log T}\left(\frac{K}{\sqrt{\nu}} + 2\sqrt{KT}\right)\right]. \tag{A.194}$$

(2) Suboptimality analysis for IRS.FH (2.61)**.** As discussed in Remark A.4.4, for IRS.FH, we have

$$Q_t^{z,\mathrm{in}}(a_t^{z,*}) - Q_t^{z,\mathrm{in}}(a) = \hat{\mu}_t^\pi(a_t^{z,*},T-t) - \hat{\mu}_t^\pi(a,T-t). \tag{A.195}$$

We construct the confidence intervals as follows:

$$U_t(a) \triangleq \hat{\mu}_t^\pi(a,0) + \sqrt{2\log T}\times\Delta_t^\pi(a,T-t), \quad L_t(a) \triangleq \hat{\mu}_t^\pi(a,0) + \sqrt{2\log T}\times\Delta_t^\pi(a,T-t). \tag{A.196}$$

Given that $\hat{\mu}_t^\pi(a,T-t)$ is $\Delta_t^\pi(a,T-t)$-sub-Gaussian conditioned on $\mathcal{F}_{t-1}$, by Lemma A.4.4, we have

$$\mathbb{E}\left[\left(\hat{\mu}_t^\pi(a,T-t) - U_t(a)\right)^+\middle|\mathcal{F}_{t-1}\right] \le \frac{\Delta_t^\pi(a,T-t)}{\sqrt{2\log T}}e^{-\frac{2\log T}{2}} \le \frac{\Delta_t^\pi(a,\infty)}{\sqrt{2\log T}}e^{-\frac{2\log T}{2}} \le \frac{\sqrt{L/\nu}}{T}. \tag{A.197}$$

Symmetrically, we have $\mathbb{E}\left[\left(L_t(a) - \hat{\mu}_t^\pi(a,T-t)\right)^+\middle|\mathcal{F}_{t-1}\right] \le \frac{\sqrt{L/\nu}}{T}$.

On the other hand, since $N_{t-1}(a) \le t$ in any case, we have

$$\frac{1}{\nu+N_{t-1}^\pi(a)}\times\frac{T-t}{\nu+N_{t-1}^\pi(a)+T-t} = \frac{1}{\nu+N_{t-1}^\pi(a)}\times\left(1-\frac{\nu+N_{t-1}^\pi(a)}{\nu+N_{t-1}^\pi(a)+T-t}\right) \tag{A.198}$$

$$= \frac{1}{\nu+N_{t-1}^\pi(a)} - \frac{1}{\nu+N_{t-1}^\pi(a)+T-t} \tag{A.199}$$

$$\le \frac{1}{\nu+N_{t-1}^\pi(a)} - \frac{1}{\nu+T}. \tag{A.200}$$

Consequently,

$$\sum_{t=1}^{T}\sqrt{\frac{1}{\nu+N_{t-1}^{\pi}(a)}-\frac{1}{\nu+T}}=\sum_{a\in\mathcal{A}}\sum_{n=0}^{N_T^{\pi}(a)-1}\sqrt{\frac{1}{\nu+n}-\frac{1}{\nu+T}} \tag{A.201}$$

$$=\sum_{a\in\mathcal{A}}\left(\sqrt{\frac{1}{\nu}-\frac{1}{\nu+T}}+\sum_{n=1}^{N_T^{\pi}(a)-1}\sqrt{\frac{1}{\nu+n}-\frac{1}{\nu+T}}\right) \tag{A.202}$$

$$\leq\frac{K}{\sqrt{\nu}}+\sum_{a\in\mathcal{A}}\sum_{n=1}^{N_T^{\pi}(a)-1}\sqrt{\frac{1}{n}-\frac{1}{T}} \tag{A.203}$$

$$\overset{(i)}{\leq}\frac{K}{\sqrt{\nu}}+\sum_{a\in\mathcal{A}}\sum_{n=1}^{N_T^{\pi}(a)-1}\left(\frac{1}{\sqrt{n}}-\frac{\sqrt{n}}{2T}\right) \tag{A.204}$$

$$\leq\frac{K}{\sqrt{\nu}}+\sum_{a\in\mathcal{A}}\int_{0}^{N_T^{\pi}(a)}\left(\frac{1}{\sqrt{x}}-\frac{\sqrt{x}}{2T}\right)dx \tag{A.205}$$

$$=\frac{K}{\sqrt{\nu}}+\sum_{a\in\mathcal{A}}\left(2\sqrt{N_T^{\pi}(a)}-\frac{\left(N_T^{\pi}(a)\right)^{3/2}}{2T}\right) \tag{A.206}$$

$$\overset{(ii)}{\leq}\frac{K}{\sqrt{\nu}}+2\sqrt{KT}-\frac{1}{3}\sqrt{T/K}, \tag{A.207}$$

where we have utilized that (i) the concavity of $\sqrt{\cdot}$, and (ii) $\min\{\sum_{a=1}^{K}n_a^{3/2};\sum_{a=1}^{K}n_a=T\}=\sum_{a=1}^{K}(T/K)^{3/2}=\sqrt{T^3/K}$.

Combining all these results, we conclude that

$$W^{\textsc{Irs.FH}}(T,\mathbf{y})-V(\pi^{\textsc{Irs.FH}},T,\mathbf{y})\leq 2\sqrt{\frac{L}{\nu}}+2\sqrt{2\log T}\sum_{t=1}^{T}\Delta_t^{\pi}(A_t^{\pi},T-t) \tag{A.208}$$

$$\leq 2\sqrt{L}\left[\frac{1}{\sqrt{\nu}}+\sqrt{2\log T}\left(\frac{K}{\sqrt{\nu}}+2\sqrt{KT}-\frac{1}{3}\sqrt{T/K}\right)\right]. \tag{A.209}$$

(3) Suboptimality analysis for IRS.V-ZERO (2.62)**.** As discussed in Remark A.4.4, for IRS.FH, we have

$$Q_t^{z,\text{in}}(a_t^{z,*})-Q_t^{z,\text{in}}(a)=\max_{0\leq n\leq T-t}\left\{\hat{\mu}_t^{\pi}(a_t^{z,*},n)\right\}-\hat{\mu}_t^{\pi}(a,0). \tag{A.210}$$

We construct the confidence intervals as follows:

$$U_t(a) \triangleq \hat{\mu}_t^\pi(a,0) + \sqrt{2\log T} \times \Delta_t^\pi(a, T-t), \quad L_t(a) \triangleq \hat{\mu}_t^\pi(a,0). \tag{A.211}$$

By Lemma A.4.6, we have

$$\mathbb{E}\left[\left(\max_{0\le n\le T-t} \hat{\mu}_t^\pi(a,n) - U_t(a)\right)^+ \middle| \mathcal{F}_{t-1}\right] \le \frac{\Delta_t^\pi(a,T-t)}{\sqrt{2\log T}} e^{-\frac{2\log T}{2}} \le \frac{\sqrt{L/\nu}}{T}, \tag{A.212}$$

where

$$\mathbb{E}\left[\hat{\mu}_t^\pi(a,0) - L_t(a) \middle| \mathcal{F}_{t-1}\right] = 0. \tag{A.213}$$

The rest of the proof is almost identical to the case of IRS.FH:

$$W^{\text{IRS.V-ZERO}}(T,\mathbf{y}) - V(\pi^{\text{IRS.V-ZERO}}, T, \mathbf{y}) \le \sqrt{\frac{L}{\nu}} + \sum_{t=1}^{T} \mathbb{E}\left[U_t(A_t^\pi) - L_t(A_t^\pi)\right]. \tag{A.214}$$

$$= \sqrt{\frac{L}{\nu}} + \sqrt{2\log T} \sum_{t=1}^{T} \Delta_t^\pi(A_t^\pi, T-t) \tag{A.215}$$

$$\le \sqrt{L}\left[\frac{1}{\sqrt{\nu}} + \sqrt{2\log T}\left(\frac{K}{\sqrt{\nu}} + 2\sqrt{KT} - \frac{1}{3}\sqrt{T/K}\right)\right]. \tag{A.216}$$

Appendix B: Appendix for Risk-sensitive Optimal Execution via a Conditional Value-at-Risk Objective

Organization of appendix. The appendix is organized as follows. In Appendix B.1, we identify the optimal deterministic strategy and its performance for §4.5.1. The CVaR performance of the optimal deterministic strategy is utilized as an upper bound on the CVaR performance of the optimal adaptive strategy. In Appendix B.2, we provide the basic characterizations of S-CVaR measure introduced in §4.2. In Appendix B.3, we provide the preliminary characterizations of the value function, and by applying Sion's minimax theorem, we prove Theorem 4.3.2 and Theorem 4.3.3 stated in §4.3. The main challenge here is to verify the conditions of Sion's minimax theorem. In Appendix B.4, we provide proofs for §4.4. We first state and prove Theorem B.4.1 from which Theorem 4.4.1 and Theorem 4.4.3 follow almost immediately. Proposition 4.4.1, Proposition 4.4.2 and Theorem 4.4.2 are proven separately.

B.1 Optimal deterministic schedules

Lemma B.1.1. *For any $a, b > 0$,*

$$\min_{x\in\mathbb{R}_+}\left\{\frac{a}{x}+b\sqrt{x}\right\} = \frac{a}{x}+b\sqrt{x}\,\Big|_{x=\left(\frac{2a}{b}\right)^{\frac{2}{3}}} = \frac{3a^{\frac{1}{3}}b^{\frac{2}{3}}}{2^{\frac{2}{3}}}. \tag{B.1}$$

Proof. Let $f(x) \triangleq \frac{a}{x}+b\sqrt{x}$. Since $f'(x) = -\frac{a}{x^2}+\frac{b}{2\sqrt{x}}$, the equation $f'(x) = 0$ has a unique solution at $x = \left(\frac{2a}{b}\right)^{\frac{2}{3}}$. □

Proof of Proposition 4.5.1. We prove the optimality of exponential schedules and identify the optimal decaying rate.

First we consider a mean-variance optimization problem:

$$\text{minimize}_{\pi\in\mathcal{D}(x)}\ \mathbb{E}[C_\infty^{x,\pi}] + \lambda \text{Var}[C_\infty^{x,\pi}], \tag{B.2}$$

where $\mathcal{D}(x)$ is the set of all deterministic policies and $\lambda \in (0,\infty)$ is a penalty for variance term. This is equivalent to an optimization over the deterministic trajectories of $(X_t)_{t\geq 0}$:

$$\inf_{X:X_0=x} \int_{t=0}^{\infty} \left(\frac{\eta}{2}\dot{X}_t^2 + \lambda\sigma^2 X_t^2\right) dt, \tag{B.3}$$

where $\dot{X}_t \triangleq \frac{d}{dt}X_t$. By applying calculus of variations, the optimal schedule $X^\star$ has to satisfy $\lambda\sigma^2 X_t^\star - \eta\ddot{X}_t^\star = 0$, at each time t, with boundary conditions $X_0^\star = x$ and $\lim_{t\to\infty} X_t^\star = 0$. The solution is uniquely given by an exponential schedule

$$X_t^\star = x\exp\left(-t/\rho_\lambda\right), \tag{B.4}$$

with the decay rate $\rho_\lambda \triangleq \sqrt{\frac{2\eta}{\lambda\sigma^2}} \in (0,\infty)$, and such a schedule yields

$$\mathbb{E}[C_\infty^{x,\pi^\star}] = \frac{\eta x^2}{4\rho_\lambda}, \quad \text{Var}[C_\infty^{x,\pi^\star}] = \frac{\sigma^2 x^2 \rho_\lambda}{2}. \tag{B.5}$$

In other words, the efficient frontier of the range of mean and variance achievable by deterministic schedules, $\left\{\left(\mathbb{E}[C_\infty^{x,\pi}], \text{Var}[C_\infty^{x,\pi}]\right)\right\}_{\pi\in\mathcal{D}(x)}$, is given by $\left\{\left(\frac{\eta x^2}{4\rho}, \frac{\sigma^2 x^2 \rho}{2}\right)\right\}_{\rho\in(0,\infty)}$ and is attained by exponential schedules.

Let us now consider the achievable range of mean and deviation, $\left\{\left(\mathbb{E}[C_\infty^{x,\pi}], \sqrt{\text{Var}[C_\infty^{x,\pi}]}\right)\right\}_{\pi\in\mathcal{D}(x)}$. Observe that its efficient frontier is still characterized by $\left\{\left(\frac{\eta x^2}{4\rho}, \sqrt{\frac{\sigma^2 x^2 \rho}{2}}\right)\right\}_{\rho\in(0,\infty)}$. Therefore, the optimal solution of the following mean-deviation optimization problem

$$\text{minimize}_{\pi\in\mathcal{D}(x)}\ \mathbb{E}[C_\infty^{x,\pi}] + \theta\sqrt{\text{Var}[C_\infty^{x,\pi}]}, \tag{B.6}$$

is also given by an exponential schedule, for any given $\theta \in (0,\infty)$.

Finally, observe that for any deterministic schedule $\pi \in \mathcal{D}(x)$ the resulting cost $C_\infty^{x,\pi}$ is normally distributed. Consequently, for any $\pi \in \mathcal{D}(x)$ and $q \in (0,1)$, we further have

$$\mathrm{CVaR}_q\left[C_\infty^{x,\pi}\right] = \mathbb{E}\left[C_\infty^{x,\pi}\right] + \frac{\kappa(q)}{q}\sqrt{\mathrm{Var}\left[C_\infty^{x,\pi}\right]}, \tag{B.7}$$

and thus $\mathrm{CVaR}_q\left[C_\infty^{x,\pi}\right]$ is minimized by an exponential schedule. Using (B.5), the optimal time constant can be determined as

$$\tau^\star \triangleq \operatorname*{argmin}_\rho \left\{\frac{\eta x^2}{4\rho} + \frac{\kappa(q)}{q}\sqrt{\frac{\sigma^2 x^2 \rho}{2}}\right\} = \left(\frac{\eta x q}{\sqrt{2}\sigma\kappa(q)}\right)^{\frac{2}{3}}. \tag{B.8}$$

This concludes the proof. □

Proof of Proposition 4.5.2. With some calculation, it can be easily shown that a TWAP schedule $X_t = x\left(1 - \frac{t}{T}\right)^+$ yields

$$\mathbb{E}[C_\infty^{x,\pi}] = \frac{\eta x^2}{2T}, \quad \mathrm{Var}[C_\infty^{x,\pi}] = \frac{\sigma^2 x^2 T}{3}. \tag{B.9}$$

Therefore, the optimal execution horizon $T^\star$ is given by

$$T^\star \triangleq \operatorname*{argmin}_T \left\{\frac{\eta x^2}{2T} + \frac{\kappa(q)}{q}\sqrt{\frac{\sigma^2 x^2 T}{3}}\right\} = \left(\frac{\sqrt{3}\eta x q}{\sigma\kappa(q)}\right)^{\frac{2}{3}}. \tag{B.10}$$

□

We state the following lemma that identifies the boundary values of $\text{S-CVaR}_q[C_\infty^{x,\mathrm{EXP}}]$, which is useful to characterize the optimal value function.

Lemma B.1.2. *The function $\kappa(q)$ given in* (4.43) *satisfies*

$$\sup_{q\in[0,1]} \kappa(q) < \infty, \quad \lim_{n\to\infty} \sqrt{n}\kappa(1/n) = \lim_{n\to\epsilon} \sqrt{n}\kappa(1-1/n) = 0. \tag{B.11}$$

Proof. Recall that $\kappa(q) \triangleq \phi\left(\Phi^{-1}(1-q)\right)$. Since $\phi(z) = \frac{1}{\sqrt{2\pi}}e^{-x^2/2} \le \frac{1}{\sqrt{2\pi}}$ for any $z \in \mathbb{R}$, we have $\sup_{q\in[0,1]} \kappa(q) \le \frac{1}{\sqrt{2\pi}} < \infty$. Also note that $\kappa(q) = \kappa(1-q)$ since $\Phi^{-1}(q) = -\Phi^{-1}(1-q)$ and

$\phi(z) = \phi(-z)$. Therefore, it suffices to show that $\lim_{n\to\infty} \sqrt{n}\kappa(1/n) = 0$.

We have the following tail bounds of standard normal distribution [75, Theorem 1.2.3]: for any $z > 0$,

$$\left(z^{-1} - z^{-3}\right) \phi(z) \leq 1 - \Phi(z) \leq z^{-1}\phi(z). \tag{B.12}$$

Define $z_n \triangleq \Phi^{-1}(1 - 1/n) > 0$, and then we have

$$\frac{\phi(z_n)}{2z_n} \leq \frac{1}{n} \leq \frac{\phi(z_n)}{z_n}, \tag{B.13}$$

for large enough n (such that $z_n^{-3} \leq \frac{1}{2} z_n^{-1}$) since $\lim_{n\to\infty} z_n = \infty$. Observe that $1/n \leq \phi(z_n)/z_n \leq \phi(z_n) = \exp(-z_n^2/2)/\sqrt{2\pi} \leq \exp(-z_n^2/2)$ and thus $z_n \leq \sqrt{2\log n}$. We further deduce that, since $\kappa(1/n) = \phi(\Phi^{-1}(1 - 1/n)) = \phi(z_n)$,

$$\sqrt{n}\kappa(1/n) = \sqrt{n}\phi(z_n) = 2\sqrt{n}z_n \times \frac{\phi(z_n)}{2z_n} \leq 2\sqrt{n}z_n \times \frac{1}{n} \leq \frac{2z_n}{\sqrt{n}} \leq 2\sqrt{\frac{2\log n}{n}}. \tag{B.14}$$

Therefore, $\lim_{n\to\infty} \sqrt{n}\kappa(1/n) = 0$. □

B.2 Preliminary characterizations of S-CVaR

Lemma B.2.1. *The risk envelope $\mathcal{Q}(q)$ is a non-empty, convex and weakly* compact subset of $\mathcal{L}^\infty$.*

Proof. It is non-empty because $Q(\omega) = q$ is always feasible.

Consider $Q_1, Q_2 \in \mathcal{Q}(q)$ and $Q_\lambda \triangleq \lambda Q_1 + (1-\lambda)Q_2$ for some $\lambda \in [0,1]$. Since $Q_1(\omega), Q_2(\omega) \in [0,1]$, we have $Q_\lambda(\omega) \in [0,1]$, and by the linearity of expectation, $\mathbb{E}[Q_\lambda] = \lambda\mathbb{E}[Q_1] + (1-\lambda)\mathbb{E}[Q_2] = q$. Therefore, $Q_\lambda \in \mathcal{Q}(q)$ and thus $\mathcal{Q}(q)$ is convex.

Finally, note that $\mathcal{Q}(q)$ is a closed subset of the unit ball in $\mathcal{L}^\infty(\Omega, \mathcal{F}, \mathbb{P})$. Given that $\mathcal{L}^\infty$ is the dual space of $\mathcal{L}^1$, it is weakly* compact by Banach-Alaoglu theorem. □

Proof of Proposition 4.2.1. *Proof of claim (i) and (v).* Claim (i) immediately follows from the dual representation of CVaR value [62, Example 6.16]:

$$\mathrm{CVaR}_q[C] = \sup_{Q \in \mathcal{L}^\infty : 0 \le Q \le \frac{1}{q}, \mathbb{E}[Q]=1} \mathbb{E}\left[CQ\right], \tag{B.15}$$

and claim (v) follows from the following identity [62, Theorem 6.2]:

$$\mathrm{CVaR}_q[C] = \mathbb{E}\left[C \,\middle|\, C \ge F_C^{-1}(1-q)\right] = \frac{\mathbb{E}\left[C\,\mathbb{I}_{\{C \ge F_C^{-1}(1-q)\}}\right]}{\mathbb{P}[C \ge F_C^{-1}(1-q)]} = \frac{\mathbb{E}\left[C\,\mathbb{I}_{\{C \ge F_C^{-1}(1-q)\}}\right]}{q}. \tag{B.16}$$

Proof of claim (ii). When $q = 0$, the risk envelope $\mathcal{Q}(q)$ has a single element $Q(\omega) = 0$, and hence, $\sup_{Q \in \mathcal{Q}(0)} \mathbb{E}[CQ] = 0$. When $q = 1$, the risk envelope $\mathcal{Q}(q)$ also has a single element $Q(\omega) = 1$, and hence, $\sup_{Q \in \mathcal{Q}(0)} \mathbb{E}[CQ] = \mathbb{E}C$.

Proof of claim (iii). For any $Q \in \mathcal{Q}(q)$, we have $|\mathbb{E}[CQ]| \le \mathbb{E}\left[|CQ|\right] \le \mathbb{E}|C|$ since $|Q| \le 1$ almost surely, and therefore, $\left|\text{S-CVaR}_q[C]\right| \le \mathbb{E}|C|$. Furthermore, $\mathcal{Q}(q)$ contains $Q(\omega) = q$, and therefore, $\mathrm{CVaR}_q[C] \ge \mathbb{E}[qC] = q\mathbb{E}C$.

Proof of claim (iv). Consider $q_1, q_2 \in [0,1]$ and $q_\lambda \triangleq \lambda q_1 + (1-\lambda)q_2$ for some $\lambda \in [0,1]$. Let $Q_1^* \in \mathrm{argmax}_{Q \in \mathcal{Q}(q_1)} \mathbb{E}[CQ]$ and $Q_2^* \in \mathrm{argmax}_{Q \in \mathcal{Q}(q_2)} \mathbb{E}[CQ]$ (the maximum is attained since $\mathcal{Q}(q)$ is weakly* compact). Let $Q_\lambda \triangleq \lambda Q_1^* + (1-\lambda)Q_2^*$. Observe that $Q_\lambda \in [0,1]$ a.s. and

$\mathbb{E}[Q_\lambda] = \lambda\mathbb{E}[Q_1^*] + (1-\lambda)\mathbb{E}[Q_2^*] = \lambda q_1 + (1-\lambda)q_2 = q_\lambda$. Therefore, $Q_\lambda \in \mathcal{Q}(q_\lambda)$. Trivially, $\mathbb{E}[CQ_\lambda] = \mathbb{E}[\lambda CQ_1^* + (1-\lambda)CQ_2^*] = \lambda\mathbb{E}[CQ_1^*] + (1-\lambda)\mathbb{E}[CQ_2^*]$, and thus

$$\text{S-CVaR}_{q_\lambda}[C] = \sup_{Q\in\mathcal{Q}(q_\lambda)} \mathbb{E}[CQ] \geq \mathbb{E}[CQ_\lambda] = \lambda\mathbb{E}[CQ_1^*] + (1-\lambda)\mathbb{E}[CQ_2^*] \tag{B.17}$$

$$= \lambda\,\text{S-CVaR}_{q_1}[C] + (1-\lambda)\,\text{S-CVaR}_{q_2}[C]. \tag{B.18}$$

This concludes the proof. □

B.3 Proofs for §4.3

From now on, we characterize S-CVaR$_q[\,\cdot\,]$ as a mapping from $\mathcal{L}^2$ to $\mathbb{R}$. This can be done without loss generality since we have $C_\infty^{x,\pi} \in \mathcal{L}^2$ for any feasible policy $\pi \in \Pi(x)$. This is to utilize the fact that $\mathcal{L}^2$ is reflexive so that its weak* topology coincides with its weak topology.

Lemma B.3.1. *The risk envelope $\mathcal{Q}(q)$ is a weakly compact subset of $\mathcal{L}^2$.*

Proof. As stated in the proof of Lemma B.2.1, $\mathcal{Q}(q)$ is a closed subset of the unit ball in $\mathcal{L}^\infty$, which is a subset of $\mathcal{L}^2$. By Banach–Alaoglu theorem, it is weakly* compact in $\mathcal{L}^2$ and hence weakly compact since $\mathcal{L}^2$ is reflexive. □

Proof of Proposition 4.3.1. *Proof of claim (i).* Note that $(Q_t^{q,\gamma})_{t\geq 0}$ is a continuous local martingale since it is a stochastic integral of a progressively measurable process with respect to Brownian motion [72, Theorem 33 in Chap. III]. Since $Q_t^{q,\gamma} \in [0,1]$ for any $t \in [0,\infty)$ by the definition of $\Gamma(q)$, it is a bounded local martingale, which is indeed a martingale [72, Thm. 51 in Chap. I].

Proof of claim (ii). The limit $\lim_{t\to\infty} Q_t^{q,\gamma}$ exists due to martingale convergence theorem [72, Theorem 10 in Chap. I].

Proof of claim (iii). Define a stopping time $\tau \triangleq \inf_{t\geq 0}\left\{Q_t^{q,\gamma} \in \{0,1\}\right\}$. Since $\left(Q_t^{q,\gamma}\right)_{t\geq 0}$ is a martingale, we have $\mathbb{E}[Q_{\tau'}^{q,\gamma}|\mathcal{F}_\tau] = Q_\tau^{q,\gamma} \in \{0,1\}$ for any $\tau' \geq \tau$. Since $0 \leq Q_{\tau'}^{q,\gamma} \leq 1$, we have $Q_{\tau'}^{q,\gamma} \in \{0,1\}$ almost surely. □

B.3.1 Proof of Theorem 4.3.2

Within this subsection, we characterize the trader's policy with its position process $(X_t)_{t\geq 0}$ rather than dealing with the liquidation rate process $(\pi_t)_{t\geq 0}$: the set of admissible policies is represented as

$$\mathcal{X}(x) \triangleq \left\{ X : \mathbb{T}\times\Omega \to \mathbb{R} \,\middle|\, X \in \mathcal{P}, X_0 = x, \mathbb{E}\left[\left(\int_{t=0}^\infty \dot{X}_t^2 dt\right)^2\right] < \infty, \mathbb{E}\left[\int_{t=0}^\infty X_t^2 dt\right] < \infty, \sup_{t\geq 0}|X_t| \leq M \right\}, \tag{B.19}$$

where $\dot{X}_t \triangleq \frac{dX_t}{dt}$ so that $\Pi(x) = \{\pi = \dot{X} | X \in \mathcal{X}(x)\}$. Accordingly, we represent the loss process as

$$C_t^X \triangleq \int_{s=0}^{t} \frac{1}{2}\eta(\dot{X}_s)^2 ds - \int_{s=0}^{t} \sigma X_s dW_s, \tag{B.20}$$

so that we have $C_t^X = C_t^{x,\pi}$ if $X \in \mathcal{X}(x)$ and $\pi = \dot{X}$.

We aim to prove Theorem 4.3.2 by utilizing Sion's minimax theorem:

Lemma B.3.2 (Sion's minimax theorem [73])**.** *Let $\mathcal{X}$ be a convex subset of a linear topological space and $\mathcal{Y}$ a compact convex subset of a linear topological space. If f is a real-valued function on $\mathcal{X} \times \mathcal{Y}$ with $f(x, \cdot)$ upper semicontinuous and quasi-concave on Y, for each $x \in \mathcal{X}$, and $f(\cdot, y)$ lower semicontinuous and quasi-convex on X, for each $y \in \mathcal{Y}$, then,*

$$\inf_{x\in\mathcal{X}} \max_{y\in\mathcal{Y}} f(x, y) = \max_{y\in\mathcal{Y}} \inf_{x\in\mathcal{X}} f(x, y). \tag{B.21}$$

Lemma B.3.3. *$\mathcal{X}(x)$ is a convex subset of a linear space endowed with a norm $\|\cdot\|_{\mathcal{X}}$:*

$$\|X\|_{\mathcal{X}} \triangleq \sqrt{\mathbb{E}\left[\left(\int_{t=0}^{\infty} \dot{X}_t^2 dt\right)^2\right]} + \sqrt{\mathbb{E}\left[\int_{t=0}^{\infty} X_t^2 dt\right]}. \tag{B.22}$$

Proof. It can be easily verified that $\|\cdot\|_{\mathcal{X}}$ is a valid norm (as a norm of Sobolev space). Also note that $\|X\|_{\mathcal{X}} < \infty$ for any $X \in \mathcal{X}(x)$. Now consider $X^{(1)}, X^{(2)} \in \mathcal{X}(x)$ and $X^{(\lambda)} \triangleq \lambda X^{(1)} + (1-\lambda)X^{(2)}$ for some $\lambda \in [0, 1]$. Observe that (i) $X_0^{(\lambda)} = x$, (ii) $\sqrt{\int_{t=0}^{\infty} \left(\dot{X}_t^{(\lambda)}\right)^2 dt} \leq \lambda\sqrt{\int_{t=0}^{\infty} \left(\dot{X}_t^{(1)}\right)^2 dt} + (1 - \lambda)\sqrt{\int_{t=0}^{\infty} \left(\dot{X}_t^{(2)}\right)^2 dt} \in \mathcal{L}^4$, and thus $\mathbb{E}\left[\left(\int_{t=0}^{\infty} \left(\dot{X}_t^{(\lambda)}\right)^2 dt\right)^2\right] < \infty$, (iii) similarly, $\mathbb{E}\left[\int_{t=0}^{\infty} \left(X_t^{(\lambda)}\right)^2 dt\right] < \infty$, and (iv) $\sup_{t\geq 0} |X_t^{(\lambda)}| \leq \lambda \sup_{t\geq 0} |X_t^{(1)}| + (1 - \lambda) \sup_{t\geq 0} |X_t^{(2)}| \leq M$. Therefore, $X^{(\lambda)} \in \mathcal{X}(x)$, and hence $\mathcal{X}(x)$ is a convex set. □

Proof of Theorem 4.3.2. By Lemma 4.3.1, it suffices to show

$$\inf_{X\in\mathcal{X}(x)} \sup_{Q\in\mathcal{Q}(q)} \mathbb{E}\left[C_\infty^X Q\right] = \sup_{Q\in\mathcal{Q}(q)} \inf_{X\in\mathcal{X}(x)} \mathbb{E}\left[C_\infty^X Q\right]. \tag{B.23}$$

We have shown that $\mathcal{X}(x)$ is a convex subset of a linear space endowed with a norm $\|\cdot\|_{\mathcal{X}}$ (Lemma

B.3.3), and $\mathcal{Q}(q)$ is a compact convex subset of $\mathcal{L}^2(\Omega, \mathcal{F}, \mathbb{P})$ endowed with the weak topology (Lemma B.3.1). In what follows, we verify the other conditions required in Lemma B.3.2:

(i) The mapping $X \mapsto \mathbb{E}[C_\infty^X Q]$ is convex and continuous in norm $\|\cdot\|_{\mathcal{X}}$ for any given $Q \in \mathcal{Q}(q)$.

(ii) The mapping $Q \mapsto \mathbb{E}[C_\infty^X Q]$ is concave and weakly continuous on $\mathcal{Q}(q)$ for any given $X \in \mathcal{X}(x)$.

Proof of claim (i). The convexity immediately follows from the fact that the mapping $X \mapsto C_\infty^X$ is quadratic. We now focus on the continuity.

Fix $X \in \mathcal{X}(x)$ and consider a sequence $\left(X^{(n)} \in \mathcal{X}(x)\right)_{n\in\mathbb{N}}$ such that $\lim_{n\to\infty} \|X - X^{(n)}\|_{\mathcal{X}} = 0$: i.e.,

$$\lim_{n\to\infty} \mathbb{E}\left[\left(\int_{t=0}^{\infty} \left(\dot{X}_t - \dot{X}_t^{(n)}\right)^2 dt\right)^2\right] = 0, \quad \lim_{n\to\infty} \mathbb{E}\left[\int_{t=0}^{\infty} \left(X_t - X_t^{(n)}\right)^2 dt\right] = 0. \tag{B.24}$$

Let

$$A \triangleq \sqrt{\int_{t=0}^{\infty} \dot{X}_t^2 dt}, \quad A_n \triangleq \sqrt{\int_{t=0}^{\infty} (\dot{X}_t^{(n)})^2 dt}, \quad \Delta_n \triangleq \sqrt{\int_{t=0}^{\infty} \left(\dot{X}_t - \dot{X}_t^{(n)}\right)^2 dt}. \tag{B.25}$$

We then have $A \in \mathcal{L}^4$, $A_n \in \mathcal{L}^4$, and $|A - A_n| \le \Delta_n$ (triangle inequality) where $\Delta_n \xrightarrow{\mathcal{L}^4} 0$ due to the above condition (B.24). Further observe that

$$\left|C_\infty^X Q - C_\infty^{X^{(n)}} Q\right| \le \left|C_\infty^X - C_\infty^{X^{(n)}}\right| \tag{B.26}$$

$$\le \frac{\eta}{2}\left|\int_{t=0}^{\infty} \dot{X}_t^2 dt - \int_{t=0}^{\infty} (\dot{X}_t^{(n)})^2 dt\right| + \sigma\left|\int_{t=0}^{\infty} X_t dW_t - \int_{t=0}^{\infty} X_t^{(n)} dW_t\right| \tag{B.27}$$

$$= \frac{\eta}{2}\left|A^2 - A_n^2\right| + \sigma\left|\int_{t=0}^{\infty} \left(X_t - X_t^{(n)}\right) dW_t\right|, \tag{B.28}$$

where the first inequality holds since $|Q| \leq 1$. Regarding the first term of (B.28), we have

$$\mathbb{E}\left[\left|A_n^2 - A^2\right|\right] = \mathbb{E}\left[\left|(A_n - A)^2 + 2A(A_n - A)\right|\right] \tag{B.29}$$

$$\leq \mathbb{E}\left[(A_n - A)^2\right] + 2\mathbb{E}\left[|A| \cdot |A_n - A|\right] \tag{B.30}$$

$$\leq \mathbb{E}\left[(A_n - A)^2\right] + 2\sqrt{\mathbb{E}\left[A^2\right]}\sqrt{\mathbb{E}\left[|A_n - A|^2\right]} \tag{B.31}$$

$$= \mathbb{E}\left[\Delta_n^2\right] + 2\sqrt{\mathbb{E}\left[A^2\right]}\sqrt{\mathbb{E}\left[\Delta_n^2\right]}. \tag{B.32}$$

Therefore, $\lim_{n\to\infty} \mathbb{E}\left[\left|A_n^2 - A^2\right|\right] = 0$ since $\Delta_n \xrightarrow{\mathcal{L}^2} 0$. Regarding the second term of (B.28), we further obtain

$$\mathbb{E}\left[\left(\int_{t=0}^{\infty}\left(X_t - X_t^{(n)}\right) dW_t\right)^2\right] = \mathbb{E}\left[\int_{t=0}^{\infty}\left(X_t - X_t^{(n)}\right)^2 dt\right], \tag{B.33}$$

which vanishes as $n \to \infty$ due to the condition (B.24). Therefore,

$$\lim_{n\to\infty} \mathbb{E}\left[\left|C_\infty^X Q - C_\infty^{X^{(n)}} Q\right|\right] = 0, \tag{B.34}$$

which implies that $\lim_{n\to\infty} \mathbb{E}\left[C_\infty^{X^{(n)}} Q\right] = \mathbb{E}\left[C_\infty^X Q\right]$. This concludes the proof.

Proof of claim (ii). The concavity immediately follows from the fact that the mapping $Q \mapsto C_\infty^X Q$ is linear. Fix $Q \in \mathcal{Q}(q)$, consider a sequence $\left(Q_n \in \mathcal{Q}(q)\right)_{n\in\mathbb{N}}$ converging to Q in $\mathcal{L}^2$-weak topology: i.e., $\lim_{n\to\infty} \mathbb{E}[ZQ_n] = \mathbb{E}[ZQ]$ for any $Z \in \mathcal{L}^2$. Since $C_\infty^X \in \mathcal{L}^2$ for any $X \in \mathcal{X}(x)$, the continuity of the mapping $Q \mapsto \mathbb{E}[C_\infty^X Q]$ immediately follows. □

B.3.2 Preliminary characterizations of value function

Proposition B.3.1. *The value function $V(\cdot,\cdot)$ satisfies the followings:*

(i) $0 \le V(x,q) \le \sigma^{\frac{2}{3}}\eta^{\frac{1}{3}} \times |x|^{\frac{4}{3}} \times q^{\frac{1}{3}}\big(\kappa(q)\big)^{\frac{2}{3}}$.

(ii) $V(x,q)$ *is convex in* x *on* $\mathbb{R}$ *for each* $q \in [0,1]$.

(iii) $V(x,q)$ *is concave in* q *on* $[0,1]$ *for each* $x \in \mathbb{R}$.

(iv) $V(x,0) = V(x,1) = V(0,q) = 0$ *for each* $x \in \mathbb{R}$ *and* $q \in [0,1]$.

Proof. Proof of claim (i). Note that $\mathbb{E}[C_\infty^{x,\pi}] \ge 0$ for any $\pi \in \Pi(x)$. By Proposition 4.2.1.(iii), we have S-CVaR$_q[C_\infty^{x,\pi}] \ge 0$ for any $\pi \in \Pi(x)$, and hence $V(x,q) \ge 0$. The upper bound follows from Proposition 4.5.1 given that $\frac{3}{2^{5/3}} < 1$.

Proof of claim (ii). Fix $q \in [0,1]$ and define $\varphi_\gamma : \mathbb{R} \mapsto \mathbb{R}$ as follows:

$$\varphi_\gamma(x) \triangleq \inf_{\pi \in \Pi(x)} J(\pi, \gamma; x, q), \tag{B.35}$$

so that we have $V(x,q) = \max_{\gamma \in \Gamma(q)} \varphi_\gamma(x)$. Since the pointwise maximum of a set of convex functions is convex, it suffices to show that $\varphi_\gamma(\cdot)$ is convex for each $\gamma \in \Gamma(q)$.

Fix $\gamma \in \Gamma(q)$ and consider any $x_1, x_2 \in \mathbb{R}$. For any $\epsilon > 0$, by definition of infimum, there exist $\pi^{1,\epsilon} \in \Pi(x_1)$ and $\pi^{2,\epsilon} \in \Pi(x_2)$ such that

$$J(\pi^{1,\epsilon}, \gamma; x_1, q) \le \varphi_\gamma(x_1) + \epsilon, \quad J(\pi^{2,\epsilon}, \gamma; x_2, q) \le \varphi_\gamma(x_2) + \epsilon. \tag{B.36}$$

Given some $\lambda \in [0,1]$, let $x_\lambda \triangleq \lambda x_1 + (1-\lambda)x_2$ and $\pi^{\lambda,\epsilon} \triangleq \lambda\pi^{1,\epsilon} + (1-\lambda)\pi^{2,\epsilon}$. Note that

$$X_t^{x_\lambda, \pi^{\lambda,\epsilon}} = x_\lambda - \int_{s=0}^{t} \pi_s^{\lambda,\epsilon} ds = \lambda X_t^{x_1, \pi^{1,\epsilon}} + (1-\lambda) X_t^{x_2 \pi^{2,\epsilon}} \tag{B.37}$$

and hence $\|\pi^{\lambda,\epsilon}\|_{\Pi(x_\lambda)} \leq \lambda\|\pi^{1,\epsilon}\|_{\Pi(x_1)} + (1-\lambda)\|\pi^{2,\epsilon}\|_{\Pi(x_2)} < \infty$, i.e., $\pi^{\lambda,\epsilon} \in \Pi(x_\lambda)$. Consequently,

$$C_\infty^{x_\lambda,\pi^{\lambda,\epsilon}} = \int_{s=0}^{t} \frac{1}{2}\eta \left|\pi_s^{\lambda,\epsilon}\right|^2 ds - \int_{s=0}^{t} \sigma X_s^{x_\lambda,\pi^{\lambda,\epsilon}} dW_s \leq \lambda C_\infty^{x_1,\pi^{1,\epsilon}} + (1-\lambda) C_\infty^{x_2,\pi^{2,\epsilon}}, \tag{B.38}$$

and therefore,

$$J(\pi^{\lambda,\epsilon},\gamma;x_\lambda,q) = \mathbb{E}\left[C_\infty^{x_\lambda,\pi^{\lambda,\epsilon}} Q_\infty^{q,\gamma}\right] \leq \lambda\mathbb{E}\left[C_\infty^{x_1,\pi^{1,\epsilon}} Q_\infty^{q,\gamma}\right] + (1-\lambda)\mathbb{E}\left[C_\infty^{x_2,\pi^{2,\epsilon}} Q_\infty^{q,\gamma}\right] \tag{B.39}$$

$$= \lambda J(\pi^{1,\epsilon},\gamma;x_1,q) + (1-\lambda) J(\pi^{2,\epsilon},\gamma;x_2,q) \tag{B.40}$$

$$\leq \lambda\varphi_\gamma(x_1) + (1-\lambda)\varphi_\gamma(x_2) + \epsilon. \tag{B.41}$$

As a result, we have

$$\varphi_\gamma(\lambda x_1 + (1-\lambda)x_2) \leq J(\pi^{\lambda,\epsilon},\gamma;x_\lambda,q) \leq \lambda\varphi_\gamma(x_1) + (1-\lambda)\varphi_\gamma(x_2) + \epsilon. \tag{B.42}$$

Since $x_1, x_2, \lambda, \epsilon$ were arbitrarily chosen, $\varphi_\gamma(\cdot)$ is convex on $\mathbb{R}$.

Proof of claim (iii). Fix $x \in \mathbb{R}$ and define $\varphi_x : [0,1] \mapsto \mathbb{R}$ as follows:

$$\varphi_\pi(q) \triangleq \sup_{\gamma\in\Gamma(q)} J(\pi,\gamma;x,q), \tag{B.43}$$

so that we have $V(x,q) = \inf_{\pi\in\Pi(x)} \varphi_\pi(q)$. Since the point-wise infimum of a set of concave functions is concave, it suffices to show that $\varphi_\pi(\cdot)$ is concave on $[0,1]$ for each $\pi \in \Pi(x)$.

Fix $\pi \in \Pi(x)$ and consider any $q_1, q_2 \in [0,1]$. Since $\Gamma(q_1)$ and $\Gamma(q_2)$ are weakly compact, there exist $\gamma^1 \in \Gamma(q_1)$ and $\gamma^2 \in \Gamma(q_2)$ such that

$$J(\pi,\gamma^1;x,q_1) = \varphi_\pi(q_1), \quad J(\pi,\gamma^2;x,q_2) = \varphi_\pi(q_2). \tag{B.44}$$

Given some $\lambda \in [0,1]$, let $q_\lambda \triangleq \lambda q_1 + (1-\lambda)q_2$ and $\gamma^\lambda \triangleq \lambda\gamma^1 + (1-\lambda)\gamma^2$ (i.e., $\gamma_t^\lambda(\omega) =$

$\lambda\gamma_t^1(\omega) + (1-\lambda)\gamma_t^2(\omega)$ for $\forall t, \omega$). Note that

$$Q_t^{q_\lambda,\gamma^\lambda} = q_\lambda + \int_{s=0}^{t} \gamma_s^\lambda dW_s = \lambda Q_t^{q_1,\gamma^1} + (1-\lambda) Q_t^{q_2,\gamma^2}, \tag{B.45}$$

and hence $Q_t^{q_\lambda,\gamma^\lambda} \in [0,1]$ almost surely for any t, i.e., $\gamma^\lambda \in \Gamma(q_\lambda)$. Therefore,

$$J(\pi,\gamma^\lambda;x,q_\lambda) = \mathbb{E}\left[C_\infty^{x,\pi} Q_\infty^{q_\lambda,\gamma^\lambda}\right] = \lambda\mathbb{E}\left[C_\infty^{x,\pi} Q_\infty^{q_1,\gamma^1}\right] + (1-\lambda)\mathbb{E}\left[C_\infty^{x,\pi} Q_\infty^{q_2,\gamma^2}\right] \tag{B.46}$$

$$= \lambda J(\pi,\gamma^1;x,q_1) + (1-\lambda) J(\pi,\gamma^2;x,q_2) \tag{B.47}$$

$$= \lambda\varphi_\pi(q_1) + (1-\lambda)\varphi_\pi(q_2). \tag{B.48}$$

As a result, we have

$$\varphi_\pi(\lambda q_1 + (1-\lambda)q_2) \geq J(\pi,\gamma^\lambda;x,q_\lambda) = \lambda\varphi_\pi(q_1) + (1-\lambda)\varphi_\pi(q_2). \tag{B.49}$$

Since q_1, q_2, λ were arbitrarily chosen, $\varphi_\pi(\cdot)$ is concave on $[0,1]$.

Proof of claim (iv). The claim immediately follows from claim (i). □

B.3.3 Proof of CVaR dynamic programming principle

We first state a proposition that is useful for proving upper-semicontinuity of a mapping with respect to the weak topology.

Lemma B.3.4 ([76], Proposition 2.10. Restated for upper-semicontinuity)**.** *Let $\mathcal{Y}$ be a locally convex space. A proper concave function $f : \mathcal{Y} \to [-\infty, \infty)$ is upper-semicontinuous on $\mathcal{Y}$ if and only if it is upper-semicontinuous with respect to the weak topology on $\mathcal{Y}$.*

We next prove a minimax theorem that includes the value function and the stopping time, which extends Theorem 4.3.2.

Proposition B.3.2. *For any stopping time $\tau : \Omega \to \mathbb{T}$, we have*

$$\inf_{\pi\in\Pi(x)} \sup_{\gamma\in\Gamma(q)} \mathbb{E}\left[C_\tau^{x,\pi} Q_\tau^{q,\gamma} + V(X_\tau^{x,\pi}, Q_\tau^{q,\gamma})\right] = \sup_{\gamma\in\Gamma(q)} \inf_{\pi\in\Pi(x)} \mathbb{E}\left[C_\tau^{x,\pi} Q_\tau^{q,\gamma} + V(X_\tau^{x,\pi}, Q_\tau^{q,\gamma})\right]. \quad \text{(B.50)}$$

Proof. Let us first define

$$\mathcal{Q}_\tau(q) \triangleq \{\mathbb{E}(Q|\mathcal{F}_\tau) : Q \in \mathcal{Q}(q)\}. \quad \text{(B.51)}$$

As analogous to Lemma 4.3.1, it can be easily shown that $\mathcal{Q}_\tau(q) = \left\{Q_\tau^{q,\gamma} | \gamma \in \Gamma(q)\right\}$. With the notation introduced in the proof of Theorem 4.3.2 (§B.3.1), it suffices to show that

$$\inf_{X\in\mathcal{X}(x)} \sup_{Q\in\mathcal{Q}_\tau} \mathbb{E}\left[C_\tau^X Q + V(X_\tau, Q)\right] = \sup_{Q\in\mathcal{Q}_\tau} \inf_{X\in\mathcal{X}(x)} \mathbb{E}\left[C_\tau^X Q + V(X_\tau, Q)\right]. \quad \text{(B.52)}$$

We aim to prove the followings:

(i) The mapping $X \mapsto \mathbb{E}\left[C_\tau^X Q + V(X_\tau, Q)\right]$ is convex and continuous on $\mathcal{X}(x)$ with respect to the norm $\|\cdot\|_{\mathcal{X}}$ for any given $Q \in \mathcal{Q}_\tau(q)$.

(ii) $\mathcal{Q}_\tau(q)$ is a non-empty, convex and weakly compact subset of $\mathcal{L}^2(\Omega, \mathcal{F}_\tau, \mathbb{P})$.

(iii) The mapping $Q \mapsto \mathbb{E}\left[V(X_\tau, Q)\right]$ is continuous on $\mathcal{Q}_\tau(q)$ endowed with $\mathcal{L}^2$-norm for any given $X \in \mathcal{X}(x)$.

(iv) The mapping $Q \mapsto \mathbb{E}\left[C_\tau^X Q + V(X_\tau, Q)\right]$ is concave and upper-semicontinuous on $\mathcal{Q}_\tau(q)$ endowed with $\mathcal{L}^2$-weak topology for any given $X \in \mathcal{X}(x)$.

Together with the convexity of $\mathcal{X}(x)$ (Lemma B.3.3), we obtain the desired claim by utilizing Sion's minimax theorem (Lemma B.3.2).

Proof of claim (i). The convexity of the mapping $X \mapsto \mathbb{E}\left[C_\tau^X Q + V(X_\tau, Q)\right]$ immediately follows from the convexity of $V(\cdot, Q)$, which was shown in Proposition B.3.1.(ii). Now fix $X \in \mathcal{X}(x)$ and consider a sequence $(X^{(n)} \in \mathcal{X}(x))_{n\in\mathbb{N}}$ such that $\|X - X^{(n)}\|_{\mathcal{X}} \to 0$ as $n \to \infty$. Analogous to the proof of Theorem 4.3.2, we can show that $C_\tau^{X^{(n)}} Q \xrightarrow{\mathcal{L}^1} C_\tau^X Q$ as $n \to \infty$, and also that $X_\tau^{(n)} \xrightarrow{\mathcal{L}^1} X_\tau$. By continuous mapping theorem, we further obtain that $V(X_\tau^{(n)}, Q) \xrightarrow{p} V(X_\tau, Q)$ as $n \to \infty$. Given that $\sup_{t\geq 0} |X_t^{(n)}| \leq M$ for all $n \in \mathbb{N}$, the sequence $\left(|V(X_\tau^{(n)}, Q)|\right)_{n\in\mathbb{N}}$ is uniformly bounded by $\sup_{q'\in[0,1]} \left\{\sigma^{\frac{2}{3}}\eta^{\frac{1}{3}} \times M^{\frac{4}{3}} \times q'^{\frac{1}{3}}\kappa(q')^{\frac{2}{3}}\right\} < \infty$, and hence uniformly integrable. Therefore, $V(X_\tau^{(n)}, Q) \xrightarrow{\mathcal{L}^1} V(X_\tau, Q)$. Combining these results, we obtain $\lim_{n\to\infty} \mathbb{E}\left[C_\tau^{X^{(n)}} Q + V(X_\tau^{(n)}, Q)\right] = \mathbb{E}\left[C_\tau^X Q + V(X_\tau, Q\right]$, which concludes the proof.

Proof of claim (ii). The proof is identical to those of Lemma B.2.1 and B.3.1 except that $\mathcal{L}^\infty(\Omega, \mathcal{F}, \mathbb{P})$ and $\mathcal{L}^2(\Omega, \mathcal{F}, \mathbb{P})$ are replaced with $\mathcal{L}^\infty(\Omega, \mathcal{F}_\tau, \mathbb{P})$ and $\mathcal{L}^2(\Omega, \mathcal{F}_\tau, \mathbb{P})$, respectively.

Proof of claim (iii). Fix $Q \in \mathcal{Q}_\tau(q)$ and consider a sequence $(Q^{(n)} \in \mathcal{Q}_\tau(q))_{n\in\mathbb{N}}$ such that $Q^{(n)} \xrightarrow{\mathcal{L}^2} Q$ as $n \to \infty$. By continuous mapping theorem, we have $V(X_\tau, Q^{(n)}) \xrightarrow{p} V(X_\tau, Q)$. Since the sequence $\left(|V(X_\tau, Q^{(n)})|\right)_{n\in\mathbb{N}}$ is uniformly integrable (uniformly bounded by $\sup_{q'\in[0,1]} \left\{\sigma^{\frac{2}{3}}\eta^{\frac{1}{3}} \times |M|^{\frac{4}{3}} \times q'^{\frac{1}{3}}\kappa(q')^{\frac{2}{3}}\right\} < \infty$), we further have $V(X_\tau, Q^{(n)}) \xrightarrow{\mathcal{L}^1} V(X_\tau, Q)$, which concludes the proof.

Proof of claim (iv). The concavity of the mapping $Q \mapsto \mathbb{E}\left[C_\tau^X Q + V(X_\tau, Q)\right]$ immediately follows from the concavity of $V(X_\tau, \cdot)$, which was shown in Proposition B.3.1.(iii). On the other hand, by combining the result of claim (iii) with Proposition B.3.4, we can show that the mapping $Q \mapsto \mathbb{E}\left[V(X_\tau, Q)\right]$ is upper-semicontinuous on $\mathcal{Q}_\tau(q)$ endowed with $\mathcal{L}^2$-weak topology. Therefore, it suffices to show that the mapping $Q \mapsto \mathbb{E}\left[C_\tau^X Q\right]$ is upper-semicontinuous with respect to $\mathcal{L}^2$-weak topology.

Fix $Q \in \mathcal{Q}_\tau(q)$ and consider a sequence $(Q^{(n)} \in \mathcal{Q}_\tau(q))_{n\in\mathbb{N}}$ such that $\mathbb{E}[ZQ^{(n)}] \to \mathbb{E}[ZQ]$ as $n \to \infty$ for any $Z \in \mathcal{L}^2$. Since $C_\tau^X \in \mathcal{L}^2$, we have $\lim_{n\to\infty} \mathbb{E}[C_\tau^X Q^{(n)}] = \mathbb{E}[C_\tau^X Q]$, which implies

the continuity of the mapping $Q \mapsto \mathbb{E}\left[C_\tau^X Q\right]$. □

Before proving Theorem 4.3.3, we introduce additional notation to describe Markovian structure of problem. We denote by $(X_s^{t,x,\pi})_{s\geq t}$ the trader's position process under control π that begins from the value x at time t: i.e.,

$$X_s^{t,x,\pi} \triangleq x - \int_{u=t}^{s} \pi_u du. \tag{B.53}$$

We define the adversary's martingale process $(Q_s^{t,q,\gamma})_{s\geq t}$ and the loss process $(C_s^{t,x,\pi})_{s\geq t}$ analogously:

$$Q_s^{t,q,\gamma} \triangleq q + \int_{u=t}^{s} \gamma_u dW_u, \quad C_s^{t,x,\pi} \triangleq \int_{u=t}^{s} \frac{1}{2}\eta\pi_u^2 du - \int_{u=t}^{s} \sigma X_u^{t,x,\pi} dW_u. \tag{B.54}$$

With this notation, we can describe the aforementioned processes in a recursive way:

$$X_s^{0,x,\pi} = X_s^{t,X_t^{0,x,\pi},\pi}, \quad Q_s^{0,q,\gamma} = Q_s^{t,Q_t^{0,q,\gamma},\gamma}, \quad C_s^{0,q,\gamma} = C_t^{0,x,\pi} + C_s^{t,X_t^{0,x,\pi},\pi}. \tag{B.55}$$

The policy spaces are defined analogously as well:

$$\Pi_t(x) \triangleq \{\pi \in \Pi(x) \mid \pi_s = 0, \forall s < t\}, \quad \Gamma_t(q) \triangleq \{\gamma \in \Gamma(q) \mid \gamma_s = 0, \forall s < t\}. \tag{B.56}$$

We prove Theorem 4.3.3 by utilizing Proposition B.3.2 and Proposition 4.3.2.

Proof of Theorem 4.3.3. Define

$$U(x,q) \triangleq \inf_{\pi\in\Pi(x)} \max_{\gamma\in\Gamma(q)} \mathbb{E}\left[C_\tau^{0,x,\pi} Q_\tau^{0,q,\gamma} + V\left(X_\tau^{0,x,\pi}, Q_\tau^{0,q,\gamma}\right)\right]. \tag{B.57}$$

We aim to prove $V(x,q) = U(x,q)$.

Proof of "$V(x,q) \leq U(x,q)$". Fix $\epsilon > 0$. By definition of $U(x,q)$, there exists $\pi^\circ \in \Pi(x)$ such that

$$U(x,q) + \epsilon \geq \max_{\gamma\in\Gamma(q)} \mathbb{E}\left[C_\tau^{0,x,\pi^\circ} Q_\tau^{0,q,\gamma} + V\left(X_\tau^{0,x,\pi^\circ}, Q_\tau^{0,q,\gamma}\right)\right]. \tag{B.58}$$

On the other hand, we have that for any $t \geq 0$, $\hat{x} \in \mathbb{X}$, and $\hat{q} \in [0,1]$,

$$V(\hat{x}, \hat{q}) = \inf_{\hat{\pi} \in \Pi_t(\hat{x})} \max_{\hat{\gamma} \in \Gamma_t(\hat{q})} \mathbb{E}\left[C_\infty^{t,\hat{x},\hat{\pi}} Q_\infty^{t,\hat{q},\hat{\gamma}}\right], \tag{B.59}$$

by time homogeneity of the problem. Consequently, for each $\omega \in \Omega$ and $\gamma \in \Gamma(q)$, there exists $\hat{\pi}^{\omega,\gamma} \in \Pi_{\tau(\omega)}(X_\tau^{0,x,\pi^\circ}(\omega))$ such that

$$V\left(X_\tau^{0,x,\pi^\circ}(\omega), Q_\tau^{0,q,\gamma}(\omega)\right) + \epsilon \geq \max_{\hat{\gamma} \in \Gamma_{\tau(\omega)}(Q_\tau^{0,q,\gamma}(\omega))} \mathbb{E}\left[C_\infty^{\tau, X_\tau^{0,x,\pi^\circ}, \hat{\pi}^{\omega,\gamma}} Q_\infty^{\tau, Q_\tau^{0,q,\gamma}, \hat{\gamma}} \middle| \mathcal{F}_\tau\right](\omega) \tag{B.60}$$

$$\geq \mathbb{E}\left[C_\infty^{\tau, X_\tau^{0,x,\pi^\circ}, \hat{\pi}^{\omega,\gamma}} Q_\infty^{\tau, Q_\tau^{0,q,\gamma}, \gamma} \middle| \mathcal{F}_\tau\right](\omega), \tag{B.61}$$

where the second inequality follows from the fact that the given γ may not be optimal for the period $t \geq \tau$.

For each $\gamma \in \Gamma(q)$, consider a trader's policy $\pi^\gamma : \mathbb{T} \times \Omega \to \mathbb{R}$ constructed as follows:

$$\pi_t^\gamma(\omega) \triangleq \begin{cases} \pi_t^\circ(\omega) & \text{if } t < \tau(\omega), \\ \hat{\pi}_t^{\omega,\gamma}(\omega) & \text{if } t \geq \tau(\omega), \end{cases} \tag{B.62}$$

i.e., it implements an ϵ-optimal solution to the Bellman equation before τ, and then implements another ϵ-optimal solution for the remaining horizon. We observe that $\pi^\gamma \in \Pi(x)$ since $\pi_t^\circ(\omega) \in$

$\Pi(x)$ and $\hat{\pi}_t^{\omega,\gamma} \in \Pi(x)$. Combining all there results, we obtain

$$U(x,q) \geq \max_{\gamma\in\Gamma(q)} \mathbb{E}\left[C_\tau^{0,x,\pi^\circ} Q_\tau^{0,q,\gamma} + V\left(X_\tau^{0,x,\pi^\circ}, Q_\tau^{0,q,\gamma}\right)\right] - \epsilon \tag{B.63}$$

$$\geq \max_{\gamma\in\Gamma(q)} \mathbb{E}\left[C_\tau^{0,x,\pi^\circ} Q_\tau^{0,q,\gamma} + \mathbb{E}\left[C_\infty^{\tau,X_\tau^{0,x,\pi^\circ},\hat{\pi}^{\omega,\gamma}} Q_\infty^{\tau,Q_\tau^{0,q,\gamma},\gamma} \middle| \mathcal{F}_\tau\right]\right] - 2\epsilon \tag{B.64}$$

$$= \max_{\gamma\in\Gamma(q)} \mathbb{E}\left[C_\tau^{0,x,\pi^\gamma} Q_\tau^{0,q,\gamma} + C_\infty^{\tau,X_\tau^{0,x,\pi^\gamma},\pi^\gamma} Q_\infty^{\tau,Q_\tau^{0,q,\gamma},\gamma}\right] - 2\epsilon \tag{B.65}$$

$$= \max_{\gamma\in\Gamma(q)} \mathbb{E}\left[C_\infty^{0,x,\pi^\gamma} Q_\infty^{0,q,\gamma}\right] - 2\epsilon \tag{B.66}$$

$$\geq \max_{\gamma\in\Gamma(q)} \inf_{\pi\in\Pi(x)} \mathbb{E}\left[C_\infty^{0,x,\pi} Q_\infty^{0,q,\gamma}\right] - 2\epsilon \tag{B.67}$$

$$= V(x,q) - 2\epsilon. \tag{B.68}$$

Since the choice of ϵ was arbitrary, we deduce that $U(x,q) \geq V(x,q)$.

Proof of "$V(x,q) \geq U(x,q)$". By minimax equality result for $U(x,q)$ (Proposition B.3.2), there exists $\gamma^\circ \in \Gamma(q)$ such that

$$U(x,q) = \inf_{\pi\in\Pi(x)} \mathbb{E}\left[C_\tau^{0,x,\pi} Q_\tau^{0,q,\gamma^\circ} + V\left(X_\tau^{0,x,\pi}, Q_\tau^{0,q,\gamma^\circ}\right)\right]. \tag{B.69}$$

By minimax equality result for $V(x,q)$ (Theorem 4.3.2) and by time homogeneity of the problem, we further have that for any $t \geq 0$, $\hat{x} \in \mathbb{X}$, and $\hat{q} \in [0,1]$,

$$V(\hat{x},\hat{q}) = \max_{\hat{\gamma}\in\Gamma_t(\hat{q})} \inf_{\hat{\pi}\in\Pi_t(\hat{x})} \mathbb{E}\left[C_\infty^{t,\hat{x},\hat{\pi}} Q_\infty^{t,\hat{q},\hat{\gamma}}\right]. \tag{B.70}$$

Therefore, for each $\omega \in \Omega$ and $\pi \in \Pi(x)$, there exists $\hat{\gamma}^{\omega,\pi} \in \Gamma_{\tau(\omega)}(Q_\tau^{0,q,\gamma^\circ}(\omega))$ such that

$$V\left(X_\tau^{0,x,\pi}(\omega), Q_\tau^{0,q,\gamma^\circ}(\omega)\right) = \inf_{\hat{\pi}\in\Pi_{\tau(\omega)}(X_\tau^{0,x,\pi}(\omega))} \mathbb{E}\left[C_\infty^{\tau,X_\tau^{0,x,\pi},\hat{\pi}} Q_\infty^{\tau,Q_\tau^{0,q,\gamma^\circ},\hat{\gamma}^{\omega,\pi}} \middle| \mathcal{F}_\tau\right](\omega) \tag{B.71}$$

$$\leq \mathbb{E}\left[C_\infty^{\tau,X_\tau^{0,x,\pi},\pi} Q_\infty^{\tau,Q_\tau^{0,q,\gamma^\circ},\hat{\gamma}^{\omega,\pi}} \middle| \mathcal{F}_\tau\right](\omega). \tag{B.72}$$

For each $\pi \in \Pi(x)$, consider an adversary's policy $\gamma^\pi : \mathbb{T} \times \Omega \to \mathbb{R}$ constructed as

$$\gamma_t^\pi(\omega) \triangleq \begin{cases} \gamma_t^\circ(\omega) & \text{if } t < \tau(\omega), \\ \hat{\gamma}_t^{\omega,\pi}(\omega) & \text{if } t \geq \tau(\omega). \end{cases} \tag{B.73}$$

We observe that $\gamma^\pi \in \Gamma(q)$. Combining all there results, we obtain

$$\begin{aligned} U(x,q) &= \inf_{\pi\in\Pi(x)} \mathbb{E}\left[C_\tau^{0,x,\pi} Q_\tau^{0,q,\gamma^\circ} + V\left(X_\tau^{0,x,\pi}, Q_\tau^{0,q,\gamma^\circ}\right)\right] && \text{(B.74)} \\ &\leq \inf_{\pi\in\Pi(x)} \mathbb{E}\left[C_\tau^{0,x,\pi} Q_\tau^{0,q,\gamma^\circ} + \mathbb{E}\left[C_\infty^{\tau,X_\tau^{0,x,\pi},\pi} Q_\infty^{\tau,Q_\tau^{0,q,\gamma^\circ},\hat{\gamma}^{\omega,\pi}} \middle| \mathcal{F}_\tau\right]\right] && \text{(B.75)} \\ &= \inf_{\pi\in\Pi(x)} \mathbb{E}\left[C_\tau^{0,x,\pi} Q_\tau^{0,q,\gamma^\pi} + C_\infty^{\tau,X_\tau^{0,x,\pi},\pi} Q_\infty^{\tau,Q_\tau^{0,q,\gamma^\pi},\gamma^\pi}\right] && \text{(B.76)} \\ &= \inf_{\pi\in\Pi(x)} \mathbb{E}\left[C_\infty^{0,X_0^{0,x,\pi},\pi} Q_\infty^{0,q,\gamma^\pi}\right] && \text{(B.77)} \\ &\leq \inf_{\pi\in\Pi(x)} \max_{\gamma\in\Gamma(q)} \mathbb{E}\left[C_\infty^{0,X_0^{0,x,\pi},\pi} Q_\infty^{0,q,\gamma}\right] = V(x,q). && \text{(B.78)} \end{aligned}$$

This concludes the proof. □

B.4 Proofs for §4.4

B.4.1 Proof of Theorem 4.4.1

Consider a function $V^\star(x,q)$ that satisfies the conditions of Theorem 4.4.1, and define

$$f^\star(x,q) \triangleq \frac{V^\star_x(x,q)}{\eta q}, \quad g^\star(x,q) \triangleq \frac{\sigma x}{V^\star_{qq}(x,q)}.$$

These functions should be well-defined on $\mathbb{R}\times(0,1)$ due to the condition (i). By the condition (iii), we have $V^\star_x(0,q)=0$, and consequently, by the condition (iv), we have $f^\star(0,q)=0$ and $f^\star(x,q)\geq 0$ for $x\geq 0$. Also observe that

$$f^\star(x,q) = \operatorname*{argmin}_{v\in\mathbb{R}}\left\{\frac{\eta}{2}qv^2 - V^\star_x(x,q)\,v\right\}, \quad g^\star(x,q) = \operatorname*{argmax}_{w\in\mathbb{R}}\left\{\frac{1}{2}V^\star_{qq}(x,q)\,w^2 - \sigma x w\right\}, \tag{B.79}$$

for any $x\in\mathbb{R}$ and $q\in(0,1)$.

We further define a sequence of function pairs $(f^{(n)},g^{(n)})_{n\in\mathbb{N}}$:

$$f^{(n)}(x,q) \triangleq \begin{cases} f^\star(x,q) & \text{if } (x,q)\in\mathcal{A}_n,\\ x/n & \text{if } (x,q)\notin\mathcal{A}_n,\end{cases} \quad g^{(n)}(x,q) \triangleq \begin{cases} g^\star(x,q) & \text{if } (x,q)\in\mathcal{A}_n,\\ 0 & \text{if } (x,q)\notin\mathcal{A}_n,\end{cases} \tag{B.80}$$

where

$$\mathcal{A}_n \triangleq \left\{(x,q)\in\mathbb{R}\times[0,1]\,\middle|\,|x|>\frac{1}{n},\ \frac{1}{n}<q<1-\frac{1}{n}\right\}. \tag{B.81}$$

We now prove the following:

Theorem B.4.1. *Fix $x>0$ and $q\in(0,1)$, and consider functions $V^\star$, $f^\star$, $g^\star$, $f^{(n)}$, and $g^{(n)}$ introduced above. For any $n>\max\{\frac{1}{x},\frac{1}{q},\frac{1}{1-q}\}$, we have the followings:*

(i) For any adversary's policy $\gamma\in\Gamma(q)$, a (X,Q)-Markov trader's policy $\pi^{(n),\gamma}$ induced by $f^{(n)}$ and coupled with γ is admissible: i.e., $\pi^{(n),\gamma}\in\Pi(x)$.

(ii) For any trader's policy $\pi\in\Pi(x)$, a (X,Q)-Markov adversary's policy $\gamma^{(n),\pi}$ induced by $g^{(n)}$

and coupled with π is admissible: i.e., $\gamma^{(n),\pi} \in \Gamma(q)$.

(iii) Mutually coupled (X, Q)-Markov policy pair $(\pi^{(n)}, \gamma^{(n)})$ induced by $(f^{(n)}, g^{(n)})$ is admissible: i.e., $\pi^{(n)} \in \Pi(x)$ and $\gamma^{(n)} \in \Gamma(q)$.

(iv) $V(x, q) \leq V^\star(x, q)$.

(v) $V(x, q) \geq V^\star(x, q)$.

(vi) For any $\gamma \in \Gamma(q)$, $(\pi^{(n),\gamma})_{n\in\mathbb{N}}$ defined in (i) satisfies $\limsup_{n\to\infty} J(\pi^{(n),\gamma}, \gamma; x, q) \leq V(x, q)$.

(vii) For any $\pi \in \Pi(x)$, $(\gamma^{(n),\pi})_{n\in\mathbb{N}}$ defined in (ii) satisfies $\liminf_{n\to\infty} J(\pi, \gamma^{(n),\pi}; x, q) \geq V(x, q)$.

(viii) $(\pi^{(n)}, \gamma^{(n)})_{n\in\mathbb{N}}$ defined in (iii) satisfies $\lim_{n\to\infty} J(\pi^{(n)}, \gamma^{(n)}; x, q) = V(x, q)$.

Proof. Throughout the proof, we define a sequence of hitting times $(\tau_n)_{n\in\mathbb{N}}$ as follows:

$$\tau_n \triangleq \inf_{t\geq 0}\{(X_t, Q_t) \notin \mathcal{A}_n\}. \tag{B.82}$$

Proof of claim (i). Under the trader's policy $\pi^{(n),\gamma}$, the position process $(X_t)_{t\geq 0}$ is described by

$$X_t = \begin{cases} x - \int_{s=0}^{t} f^\star(X_s, Q_s)\, ds, & \forall t \leq \tau_n, \\ X_{\tau_n} e^{-(t-\tau_n)/n}, & \forall t \geq \tau_n. \end{cases} \tag{B.83}$$

Observe that $X_t \in [0, x]$ for any $t \geq 0$, and in particular, X_t is monotonically decreasing over time. Given that $(Q_t)_{t\geq 0}$ has a continuous sample path and n is large enough such that $q \in \left[\frac{1}{n}, 1 - \frac{1}{n}\right]$, we also have $Q_t \in \left[\frac{1}{n}, 1 - \frac{1}{n}\right]$ for any $t \in [0, \tau_n]$. Then, the sample path of $(X_t)_{t\geq 0}$ for each ω is uniquely determined (i.e., the associated SDE has a unique strong solution): the uniqueness of sample path on $[0, \tau_n)$ follows from the fact that $f^\star(\cdot, \cdot)$ is Lipschitz continuous on $\left(\frac{1}{n}, x\right] \times \left(\frac{1}{n}, 1 - \frac{1}{n}\right) \subset \mathcal{A}_n$ due to the condition (iv), and the uniqueness on $[\tau_n, \infty)$ immediately follows from the fact that $X_t = X_{\tau_n} e^{-(t-\tau_n)/n}$ for $t \geq \tau_n$.

We then show that τ_n is almost surely bounded. For any $t < \tau_n$ (so that $(X_t, Q_t) \in \mathcal{A}_n$), by the condition (iv), we have

$$f^{(n)}(X_t, Q_t) \geq f^\star(X_t, 1 - 1/n) \geq f^\star(1/n, 1 - 1/n) := \alpha_n > 0. \tag{B.84}$$

Then, for any $t < \tau_n$, we have $dX_t/dt = -f^{(n)}(X_t, Q_t) \leq -\alpha_n$, and hence $X_{\tau_n} \leq x - \alpha_n \tau_n$. Together with the fact that $X_{\tau_n} \geq 0$, we deduce that

$$\tau_n \leq \frac{x}{\alpha_n}, \tag{B.85}$$

on any sample path.

We further have that, by the condition (iv), for any $t < \tau_n$,

$$f^{(n)}(X_t, Q_t) \leq f^\star(X_t, 1/n) \leq f^\star(x, 1/n) := \beta_n < \infty. \tag{B.86}$$

Then,

$$\int_{t=0}^{\tau_n} \pi_t^2 dt \leq \int_{t=0}^{\tau_n} \left(f^\star(X_t, Q_t)\right)^2 dt \leq \tau_n \beta_n^2 \leq \frac{x\beta_n^2}{\alpha_n}, \quad \int_{t=0}^{\tau_n} X_t^2 dt \leq \tau_n x^2 \leq \frac{x^3}{\alpha_n}, \tag{B.87}$$

on any sample path. Consequently, since $\pi_t = X_t/n$ for any $t \geq \tau_n$,

$$\int_{t=\tau_n}^{\infty} \pi_t^2 dt = \frac{1}{n^2} \int_{t=\tau_n}^{\infty} X_t^2 dt = \frac{1}{n^2} \int_{t=\tau_n}^{\infty} X_{\tau_n}^2 e^{-2(t-\tau_n)/n} dt \leq \frac{x^2}{2n}, \quad \int_{t=\tau_n}^{\infty} X_t^2 dt \leq \frac{nx^2}{2}. \tag{B.88}$$

As a result,

$$\mathbb{E}\left[\left(\int_{t=0}^{\infty} \pi_t^2 dt\right)^2\right] \leq \left(\frac{x\beta_n^2}{\alpha_n} + \frac{x^2}{2n}\right)^2 < \infty, \quad \mathbb{E}\left[\int_{t=0}^{\infty} X_t^2 dt\right] \leq \frac{x^3}{\alpha_n} + \frac{nx^2}{2} < \infty. \tag{B.89}$$

Also note that $\sup_{t\geq 0} |X_t| \leq x \leq M$. Therefore, $\pi^{(n),\gamma}$ is admissible: i.e., $\pi^{(n),\gamma} \in \Pi(x)$.

Proof of claim (ii). Under the adversary's policy $\gamma^{(n),\pi}$, the martingale process $(Q_t)_{t\geq 0}$ is described

by

$$Q_t = \begin{cases} q + \int_{s=0}^{t} g^\star\left(X_s, Q_s\right) dW_s, & \forall t \leq \tau_n, \\ Q_{\tau_n}, & \forall t \geq \tau_n. \end{cases} \tag{B.90}$$

From the condition (v), we can show that $g^\star(\cdot,\cdot)$ is Lipschitz continuous and bounded on $\left(\frac{1}{n}, \infty\right) \times \left(\frac{1}{n}, 1-\frac{1}{n}\right) \subset \mathcal{A}_n$. Therefore, for each ω, the sample path of $(Q_t)_{t\geq 0}$ is unique and continuous (given that n is large enough so that $q \in \left[\frac{1}{n}, 1-\frac{1}{n}\right]$, and $(X_t)_{t\geq 0}$ is continuous), and hence $Q_t \in \left[\frac{1}{n}, 1-\frac{1}{n}\right] \subset [0,1]$ almost surely for any $t \in [0,\infty)$, and hence $\gamma^{(n),\pi}$ is admissible.

Proof of claim (iii). Under the mutually coupled policy pair $(\pi^{(n)}, \gamma^{(n)})$, the process pair $(X_t, Q_t)_{t\geq 0}$ satisfies (B.83) and (B.90) simultaneously. By the same argument above (claim (ii)), we can show that $\gamma^{(n)}$ is admissible, and also that $\pi^{(n)}$ is admissible by claim (i).

Proof of claim (iv). Fix $\gamma \in \Gamma(q)$ and consider $\pi := \pi^{(n),\gamma} \in \Pi(x)$. For any $t < \tau_n$, since $f^{(n)}(X_t, Q_t) = f^\star(X_t, Q_t)$, by the condition (i), we have

$$\left(\frac{\eta}{2} Q_t \pi_t^2 - V_x^\star\left(X_t, Q_t\right) \pi_t\right) + \left(\frac{1}{2} V_{qq}^\star\left(X_t, Q_t\right) \gamma_t^2 - \sigma X_t \gamma_t\right) \tag{B.91}$$

$$= \left(\frac{\eta}{2} Q_t \left(f^\star\left(X_t, Q_t\right)\right)^2 - V_x^\star\left(X_t, Q_t\right) f^\star\left(X_t, Q_t\right)\right) + \left(\frac{1}{2} V_{qq}^\star\left(X_t, Q_t\right) \gamma_t^2 - \sigma X_t \gamma_t\right) \tag{B.92}$$

$$= \min_{v\in\mathbb{R}} \left\{\frac{\eta}{2} Q_t v^2 - V_x^\star\left(X_t, Q_t\right) v\right\} + \left(\frac{1}{2} V_{qq}^\star\left(X_t, Q_t\right) \gamma_t^2 - \sigma X_t \gamma_t\right) \tag{B.93}$$

$$\leq \min_{v\in\mathbb{R}} \left\{\frac{\eta}{2} Q_t v^2 - V_x^\star\left(X_t, Q_t\right) v\right\} + \max_{w\in\mathbb{R}} \left\{\frac{1}{2} V_{qq}^\star\left(X_t, Q_t\right) w^2 - \sigma X_t w\right\} \tag{B.94}$$

$$= 0. \tag{B.95}$$

By plugging this result into Proposition 4.4.1, we obtain

$$\mathbb{E}\left[C_{\tau_n} Q_{\tau_n}\right] \tag{B.96}$$

$$\leq \mathbb{E}\left[C_{\tau_n} Q_{\tau_n} + V^\star\left(X_{\tau_n}, Q_{\tau_n}\right)\right] \tag{B.97}$$

$$= V^\star(x,q) + \mathbb{E}\left[\int_{t=0}^{\tau_n} \left\{\left(\frac{\eta}{2} Q_t \pi_t^2 - V_x^\star\left(X_t, Q_t\right) \pi_t\right) + \left(\frac{1}{2} V_{qq}^\star\left(X_t, Q_t\right) \gamma_t^2 - \sigma X_t \gamma_t\right)\right\} dt\right] \tag{B.98}$$

$$\leq V^\star(x,q), \tag{B.99}$$

where the first inequality follows from the non-negativity of $V^\star(\cdot,\cdot)$. On the other hand, by the definition of τ_n and the continuity of sample paths, we have either $X_{\tau_n} = \frac{1}{n}$ or $Q_{\tau_n} \in \{\frac{1}{n}, 1-\frac{1}{n}\}$, and therefore,

$$\mathbb{E}\left[V\left(X_{\tau_n}, Q_{\tau_n}\right)\right] \leq \mathbb{E}\left[\max\left\{V(1/n, Q_{\tau_n}),\ V(X_{\tau_n}, 1/n),\ V(X_{\tau_n}, 1-1/n)\right\}\right] \tag{B.100}$$

$$\leq \sup_{q'\in[0,1]}\{V(1/n, q')\} + V(x, 1/n) + V(x, 1-1/n) := A_n(x), \tag{B.101}$$

where the second inequality follows from that $X_\tau \leq x$ under policy $\pi^{(n),\gamma}$ and $V(\cdot, q)$ is increasing on $[0,\infty)$. It can be shown easily that $\lim_{n\to\infty} A_n(x) = 0$ for any x due to Proposition B.3.1.

Since γ was chosen arbitrary, from (B.96), we deduce that

$$\sup_{\gamma\in\Gamma(q)} \mathbb{E}\left[C_{\tau_n}^{x,\pi^{(n),\gamma}} Q_{\tau_n}^{q,\gamma}\right] \leq V^\star(x,q). \tag{B.102}$$

Utilizing Theorem 4.3.3 with (B.101) and (B.102), we further obtain

$$V(x,q) \overset{\text{Thm 4.3.3}}{=} \sup_{\gamma\in\Gamma(q)} \inf_{\pi\in\Pi(x)} \mathbb{E}\left[C_{\tau_n}^{x,\pi} Q_{\tau_n}^{q,\gamma} + V\left(X_{\tau_n}^{x,\pi}, Q_{\tau_n}^{q,\gamma}\right)\right] \tag{B.103}$$

$$\leq \sup_{\gamma\in\Gamma(q)} \mathbb{E}\left[C_{\tau_n}^{x,\pi^{(n),\gamma}} Q_{\tau_n}^{q,\gamma} + V\left(X_{\tau_n}^{x,\pi^{(n),\gamma}}, Q_{\tau_n}^{q,\gamma}\right)\right] \tag{B.104}$$

$$\overset{\text{(B.101)}}{\leq} \sup_{\gamma\in\Gamma(q)} \mathbb{E}\left[C_{\tau_n}^{x,\pi^{(n),\gamma}} Q_{\tau_n}^{q,\gamma}\right] + A_n(x) \tag{B.105}$$

$$\overset{\text{(B.102)}}{\leq} V^\star(x,q) + A_n(x). \tag{B.106}$$

Since $\lim_{n\to\infty} A_n(x) = 0$, we obtain $V(x,q) \leq V^\star(x,q)$, which concludes the proof.

Proof of claim (v). The proof is almost symmetric to that of claim (iv). Fix $\pi \in \Pi(x)$, and consider

$\gamma := \gamma^{(n),\pi}$. For any $t < \tau_n$, since $g^{(n)}(X_t, Q_t) = g^\star(X_t, Q_t)$, by the condition (i), we have

$$\left(\frac{\eta}{2} Q_t \pi_t^2 - V_x^\star\left(X_t, Q_t\right) \pi_t\right) + \left(\frac{1}{2} V_{qq}^\star\left(X_t, Q_t\right) \gamma_t^2 - \sigma X_t \gamma_t\right) \tag{B.107}$$

$$= \left(\frac{\eta}{2} Q_t \pi_t^2 - V_x^\star\left(X_t, Q_t\right) \pi_t\right) + \max_{w \in \mathbb{R}} \left\{\frac{1}{2} V_{qq}^\star\left(X_t, Q_t\right) w^2 - \sigma X_t w\right\} \tag{B.108}$$

$$\geq \min_{v \in \mathbb{R}} \left\{\frac{\eta}{2} Q_t v^2 - V_x^\star\left(X_t, Q_t\right) v\right\} + \max_{w \in \mathbb{R}} \left\{\frac{1}{2} V_{qq}^\star\left(X_t, Q_t\right) w^2 - \sigma X_t w\right\} \tag{B.109}$$

$$= 0. \tag{B.110}$$

Consequently, by Proposition 4.4.1,

$$\mathbb{E}\left[C_{\tau_n} Q_{\tau_n} + V^\star\left(X_{\tau_n}, Q_{\tau_n}\right)\right] \tag{B.111}$$

$$= V^\star(x, q) + \mathbb{E}\left[\int_{t=0}^{\tau_n} \left\{\left(\frac{\eta}{2} Q_t \pi_t^2 - V_x^\star\left(X_t, Q_t\right) \pi_t\right) + \left(\frac{1}{2} V_{qq}^\star\left(X_t, Q_t\right) \gamma_t^2 - \sigma X_t \gamma_t\right)\right\} dt\right] \tag{B.112}$$

$$\geq V^\star(x, q). \tag{B.113}$$

By the definition of τ_n and the continuity of sample paths, we have either $X_{\tau_n} = \frac{1}{n}$ or $Q_{\tau_n} \in \{\frac{1}{n}, 1 - \frac{1}{n}\}$, and hence,

$$\mathbb{E}\left[V^\star\left(X_{\tau_n}, Q_{\tau_n}\right)\right] \leq \mathbb{E}\left[\max\left\{V^\star(1/n, Q_{\tau_n}),\ V^\star(X_{\tau_n}, 1/n),\ V^\star(X_{\tau_n}, 1 - 1/n)\right\}\right] \tag{B.114}$$

$$\leq \sup_{q' \in [0,1]} \{V^\star(1/n, q')\} + V^\star(M, 1/n) + V^\star(M, 1 - 1/n) := B_n, \tag{B.115}$$

where the last inequality follows from the feasibility of π (i.e., $|X_{\tau_n}| \leq M$), the monotonicity of $V^\star(\cdot, q)$ (condition (iii)), and the concavity of $V^\star(x, \cdot)$ (condition (i)). In particular, by the conditions (ii)–(i), we have $\lim_{n \to 0} B_n = 0$.

Since π was chosen arbitrary, we further deduce that

$$V^\star(x, q) \leq \inf_{\pi \in \Pi(x)} \mathbb{E}\left[C_\tau^{x,\pi} Q_{\tau_n}^{q,\gamma^{(n),\pi}} + V^\star\left(X_{\tau_n}^{x,\pi}, Q_{\tau_n}^{q,\gamma^{(n),\pi}}\right)\right] \leq \inf_{\pi \in \Pi(x)} \mathbb{E}\left[C_{\tau_n}^{x,\pi} Q_{\tau_n}^{q,\gamma^{(n),\pi}}\right] + B_n. \tag{B.116}$$

By utilizing Theorem 4.3.3 and above results, we obtain

$$V(x,q) = \inf_{\pi\in\Pi(x)} \sup_{\gamma\in\Gamma(q)} \mathbb{E}\left[C_{\tau_n}^{x,\pi} Q_{\tau_n}^{q,\gamma} + V\left(X_{\tau_n}^{x,\pi}, Q_{\tau_n}^{q,\gamma}\right)\right] \tag{B.117}$$

$$\geq \inf_{\pi\in\Pi(x)} \mathbb{E}\left[C_{\tau_n}^{x,\pi} Q_{\tau_n}^{q,\gamma^{(n),\pi}} + V\left(X_{\tau_n}^{x,\pi}, Q_{\tau_n}^{q,\gamma^{(n),\pi}}\right)\right] \tag{B.118}$$

$$\geq \inf_{\pi\in\Pi(x)} \mathbb{E}\left[C_{\tau_n}^{x,\pi} Q_{\tau_n}^{q,\gamma^{(n),\pi}}\right] \tag{B.119}$$

$$\geq V^\star(x,q) - B_n. \tag{B.120}$$

By taking $n \nearrow \infty$, we obtain the desired result.

Proof of claim (vi). To simplify notation, let $\pi := \pi^{(n),\gamma}$ and $\tau := \tau_n$ temporarily. By Proposition 4.3.2, we have

$$J(\pi,\gamma;x,q) = \mathbb{E}\left[C_\tau^{0,x,\pi} Q_\tau^{0,q,\gamma} + C_\infty^{\tau,X_\tau^{0,x,\pi},\pi} Q_\infty^{\tau,Q_\tau^{0,q,\gamma},\gamma}\right]. \tag{B.121}$$

(Recall that $C_\infty^{\tau,X_\tau^{0,x,\pi},\pi} = C_\infty^{0,x,\pi} - C_\tau^{0,x,\pi}$.) In the proof of claim (iv), in (B.96), we have shown that

$$\mathbb{E}\left[C_\tau^{0,x,\pi} Q_\tau^{0,q,\gamma}\right] \leq V^\star(x,q) = V(x,q). \tag{B.122}$$

Therefore,

$$J(\pi,\gamma;x,q) \leq V(x,q) + \mathbb{E}\left[C_\infty^{\tau,X_\tau^{0,x,\pi},\pi} Q_\infty^{\tau,Q_\tau^{0,q,\gamma},\gamma}\right] \tag{B.123}$$

$$\leq V(x,q) + \mathbb{E}\left[\text{S-CVaR}_{Q_\tau^{0,q,\gamma}}\left[C_\infty^{\tau,X_\tau^{0,x,\pi},\pi}\right]\right]. \tag{B.124}$$

To obtain the desired result, it suffices to show that $\lim_{n\to\infty} \mathbb{E}\left[\text{S-CVaR}_{Q_\tau^{0,q,\gamma}}\left[C_\infty^{\tau,X_\tau^{0,x,\pi},\pi}\right]\right] = 0$.

Note that after time τ_n, the policy $\pi^{(n),\gamma}$ trades according to the exponential schedule (i.e,. $X_t = X_{\tau_n} e^{-(t-\tau_n)/n}$), and thus $C_\infty^{\tau_n,X_{\tau_n},\pi}$ is normally distributed conditional on X_{τ_n}: more specifically, as derived in (B.5), we have

$$\mathbb{E}\left[C_\infty^{\tau_n,X_{\tau_n},\pi}\middle| X_{\tau_n}\right] = \frac{\eta X_{\tau_n}^2}{4n}, \quad \text{Var}\left[C_\infty^{\tau_n,X_{\tau_n},\pi}\middle| X_{\tau_n}\right] = \frac{n\sigma^2 X_{\tau_n}^2}{2}. \tag{B.125}$$

Consequently,

$$\text{S-CVaR}_{Q_{\tau_n}}\left[C_\infty^{\tau_n,X_{\tau_n},\pi}\right] = \frac{\eta X_{\tau_n}^2 Q_{\tau_n}}{4n} + \kappa(Q_{\tau_n}) \times \sqrt{\frac{n\sigma^2 X_{\tau_n}^2}{2}}. \tag{B.126}$$

By the definition of τ_n, we have either $X_{\tau_n} = \frac{1}{n}$ or $Q_{\tau_n} \in \left\{\frac{1}{n}, 1 - \frac{1}{n}\right\}$, and thus

$$\text{S-CVaR}_{Q_{\tau_n}}\left[C_\infty^{\tau_n,X_{\tau_n},\pi}\right] \leq \max\left\{\frac{\eta Q_{\tau_n}}{4n^3} + \kappa(Q_{\tau_n}) \times \sqrt{\frac{\sigma^2}{2n}},\ \frac{\eta X_{\tau_n}^2 Q_\tau}{4n} + \max\left\{\kappa(1/n), \kappa(1-1/n)\right\} \times \sqrt{\frac{n\sigma^2 X_{\tau_n}^2}{2}}\right\} \tag{B.127}$$

$$\leq \max\left\{\frac{\eta}{4n^3} + \sup_{q'}\{\kappa(q')\} \times \sqrt{\frac{\sigma^2}{2n}},\ \frac{\eta x^2}{4n} + \max\left\{\kappa(1/n), \kappa(1-1/n)\right\} \times \sqrt{\frac{n\sigma^2 x^2}{2}}\right\}. \tag{B.128}$$

Since $\sup_{q'} \kappa(q') < \infty$, $\lim_{n\to\infty} \sqrt{n}\kappa(1/n) = 0$, and $\lim_{n\to\infty} \sqrt{n}\kappa(1-1/n) = 0$ (Lemma B.1.2), the last term of (B.128) vanishes as $n \to \infty$. Therefore, we have $\lim_{n\to\infty} \mathbb{E}\left[\text{S-CVaR}_{Q_{\tau_n}}\left[C_\infty^{\tau_n,X_{\tau_n},\pi}\right]\right] = 0$ and this concludes the proof.

Proof of claim (vii). In the proof of claim (v), we have shown that

$$\mathbb{E}\left[C_{\tau_n}^{x,\pi} Q_{\tau_n}^{q,\gamma^{(n)},\pi} + V^\star\left(X_{\tau_n}^{x,\pi}, Q_{\tau_n}^{q,\gamma^{(n)},\pi}\right)\right] \geq V^\star(x,q) = V(x,q), \tag{B.129}$$

and $\lim_{n\to\infty} \mathbb{E}\left[V^\star\left(X_{\tau_n}^{x,\pi}, Q_{\tau_n}^{q,\gamma^{(n)},\pi}\right)\right] = 0$.

Observe that we have $Q_\infty = Q_{\tau_n}$ since $g^\star(X_t, Q_t) = 0$ for all $t \geq \tau_n$ under $\gamma^{(n),\pi}$. Therefore,

$$\mathbb{E}\left[C_\infty^{x,\pi} Q_\infty^{q,\gamma^{(n),\pi}}\right] - \mathbb{E}\left[C_{\tau_n}^{x,\pi} Q_{\tau_n}^{q,\gamma^{(n),\pi}}\right] = \mathbb{E}\left[\left(C_\infty - C_{\tau_n}\right) Q_{\tau_n}\right] \tag{B.130}$$

$$= \mathbb{E}\left[\mathbb{E}\left(\int_{t=\tau_n}^\infty \frac{\eta}{2}\pi_t^2 dt - \int_{t=\tau_n}^\infty \sigma X_t dW_t \middle| \mathcal{F}_{\tau_n}\right) Q_{\tau_n}\right] \tag{B.131}$$

$$= \mathbb{E}\left[\left(\int_{t=\tau_n}^\infty \frac{\eta}{2}\pi_t^2 dt\right) Q_{\tau_n}^{q,\gamma^{(n)}}\right] \tag{B.132}$$

$$\geq 0. \tag{B.133}$$

Consequently,

$$J(\pi, \gamma^{(n),\pi}; x, q) = \mathbb{E}\left[C_\infty^{x,\pi} Q_\infty^{q,\gamma^{(n),\pi}}\right] \geq \mathbb{E}\left[C_{\tau_n}^{x,\pi} Q_{\tau_n}^{q,\gamma^{(n),\pi}}\right] \geq V(x,q) - \mathbb{E}\left[V^\star\left(X_{\tau_n}^{x,\pi}, Q_{\tau_n}^{q,\gamma^{(n),\pi}}\right)\right]. \tag{B.134}$$

By taking $\liminf_{n\to\infty}$ on both sides, we obtain the desired result. □

Proof of Theorem 4.4.1. We aim to show that $V(x,q) = V^\star(x,q)$ for any $x \in \mathbb{R}$ and $q \in [0,1]$. From Theorem B.4.1.(iv) and (v), we deduce that $V(x,q) = V^\star(x,q)$ for any $x > 0$ and $q \in (0,1)$. By symmetry, the same argument holds for any $x < 0$ and $q \in (0,1)$. From Proposition B.3.1.(iv) and the condition (ii), we can also verify that their boundary values match: i.e., $V^\star(x,q) = V(x,q) = 0$ if $x = 0$ or $q \in \{0,1\}$. □

B.4.2 Other proofs

Proof of Proposition 4.4.1. Observe that $\widehat{V}(X_t, Q_t)$ is differentiable along the sample path before τ since $Q_t \in (0,1)$ for any $t \leq \tau$. By applying Itô's formula, we obtain

$$\widehat{V}(X_\tau, Q_\tau) = \widehat{V}(x,q) - \int_{t=0}^{\tau} \widehat{V}_x(X_t, Q_t)\pi_t dt + \int_{t=0}^{\tau} \widehat{V}_q(X_t, Q_t)\gamma_t dW_t + \frac{1}{2}\int_{t=0}^{\tau} \widehat{V}_{qq}(X_t, Q_t)\gamma_t^2 dt. \tag{B.135}$$

Recall that $C_t \triangleq \int_{t=0}^{\tau} \frac{\eta}{2}\pi_t^2 dt - \int_{t=0}^{\tau} \sigma X_t dW_t$. By applying Itô's product rule, we further obtain

$$C_\tau Q_\tau = Q_0 C_0 + \int_{t=0}^{\tau} \left(\frac{\eta}{2}\pi_t^2 Q_t - \sigma X_t \gamma_t\right) dt + \int_{t=0}^{\tau} \left(-\sigma X_t Q_t + X_t \gamma_t\right) dW_t. \tag{B.136}$$

Since $\int_{t=0}^{\tau} \widehat{V}_q(X_t, Q_t)\gamma_t dW_t$ and $\int_{t=0}^{\tau} \left(-\sigma X_t Q_t + X_t \gamma_t\right) dW_t$ are local martingales, from above results, we deduce that

$$\mathbb{E}\left[C_\tau Q_\tau + \widehat{V}(X_\tau, Q_\tau)\right] = \widehat{V}(x,q) + \mathbb{E}\left[\int_{t=0}^{\tau} \left\{\frac{\eta}{2}\pi_t^2 Q_t - \widehat{V}_x(X_t, Q_t)\pi_t + \frac{1}{2}\widehat{V}_{qq}(X_t, Q_t)\gamma_t^2 - \sigma X_t \gamma_t\right\} dt\right]. \tag{B.137}$$

We obtain the claim by rearranging terms. □

Proof of Theorem 4.4.2. We prove that the function $V^\star$, defined in (4.29), satisfies the conditions (i)–(v) in Theorem 4.4.1.

Verification of condition (i). Observe that $\varphi''(q) = -\frac{q}{\varphi^2(q)} < 0$ for any $q \in (0,1)$, and thus and $V^\star_{qq}(x,q) < 0$. The minimum/maximum in (4.27) are attainable, and we have

$$\min_{v\in\mathbb{R}} \left\{ \frac{\eta}{2} q v^2 - V^\star_x(x,q)\, v \right\} + \max_{w\in\mathbb{R}} \left\{ \frac{1}{2} V^\star_{qq}(x,q)\, w^2 - \sigma x w \right\} = -\frac{\left(V^\star_x(x,q)\right)^2}{2\eta q} - \frac{\sigma^2 x^2}{2V^\star_{qq}(x,q)} \quad \text{(B.138)}$$

Therefore, it suffices show that $V^{\star 2}_x \times V^\star_{qq} = -\sigma^2\eta \times x^2 q$:

$$\begin{aligned} V^{\star 2}_x \times V^\star_{qq} &= \left((3/4)^{\frac{2}{3}} \times \sigma^{\frac{2}{3}} \eta^{\frac{1}{3}} \times \frac{4}{3} x^{\frac{1}{3}} \times \varphi(q) \right)^2 \times \left((3/4)^{\frac{2}{3}} \times \sigma^{\frac{2}{3}} \eta^{\frac{1}{3}} \times |x|^{\frac{4}{3}} \times \varphi''(q) \right) \\ &= \sigma^2\eta \times x^2 \times \varphi^2(q) \times \varphi''(q) = -\sigma^2\eta x^2 q. \end{aligned}$$

Verification of conditions (ii)–(iii). These conditions can be easily verified by inspection.

Verification of condition (iv). Note that $\frac{V^\star_x(x,q)}{q} = (3/4)^{-\frac{1}{3}} \times \sigma^{\frac{2}{3}} \eta^{\frac{1}{3}} \times x^{\frac{1}{3}} \times \frac{\varphi(q)}{q}$. Its continuity and monotonicity (with respect to x) can be verified immediately, and it suffices to show that $\frac{\varphi(q)}{q}$ is decreasing in q.

Observe that for any q_1, q_2 such that $0 < q_1 \le q_2 \le 1$, we have $\varphi(q_1) \ge \frac{q_1}{q_2}\varphi(q_2) + \left(1 - \frac{q_1}{q_2}\right)\varphi(0) = q_1 \times \frac{\varphi(q_2)}{q_2}$ due to its concavity, and therefore $\frac{\varphi(q_1)}{q_1} \ge \frac{\varphi(q_2)}{q_2}$. As a result, $\frac{\varphi(q)}{q}$ is monotonically decreasing in q on $(0,1)$.

Verification of condition (v). Note that $\frac{x}{V^\star_{qq}(x,q)} = (3/4)^{-\frac{2}{3}} \times \sigma^{-\frac{2}{3}} \eta^{-\frac{1}{3}} \times x^{-\frac{1}{3}} \times \frac{\varphi^2(q)}{q}$. The continuity and monotonicity (with respect to x) is trivial. □

Proof of Theorem 4.4.3. The claims follow from Theorem B.4.1.(i), (ii), (iii), (vi), (vii), and (viii).

□

Proof of Proposition 4.4.2. Emden–Fowler equation deals with a differential equation with a form of $\frac{d^2\varphi}{dq^2} = Aq^n\varphi^m$, and the differential equation that we have corresponds to the case of $A = -1$, $n = 1$, and $m = -1/2$. In this case, the solutions can be written in parametric form [74, p. 2.3.27]:

for constants a and b such that $A = -\frac{9}{2}(b/a)^3$,

$$q(\theta) = a\theta^{-\frac{2}{3}}\left[\left(\theta Z'(\theta) + \frac{1}{3}Z(\theta)\right)^2 - \theta^2 Z^2(\theta)\right],\ \varphi(\theta) = b\theta^{\frac{2}{3}}Z^2(\theta),\ Z(\theta) = C_1 I_{1/3}(\theta) + C_2 K_{1/3}(\theta), \tag{B.139}$$

or

$$q(\theta) = a\theta^{-\frac{2}{3}}\left[\left(\theta Z'(\theta) + \frac{1}{3}Z(\theta)\right)^2 + \theta^2 Z^2(\theta)\right],\ \varphi(\theta) = b\theta^{\frac{2}{3}}Z^2(\theta),\ Z(\theta) = C_1 J_{1/3}(\theta) + C_2 Y_{1/3}(\theta), \tag{B.140}$$

where $\theta \in \mathbb{R}_+$ is the parameter, and C_1 and C_2 are arbitrary constants.

Note that the expression (4.34) is obtained by taking $C_1 = -\frac{2}{\pi}$ and $C_2 = 0$ in (B.139), and the expression (4.35) is obtained by taking $C_1 = \sqrt{3}$ and $C_2 = -1$ in (B.140). Therefore it suffices to show that the curve $\{(q_L(\theta), \varphi_L(\theta))\}_{\theta\in(0,\infty]} \bigcup \{(q_R(\theta), \varphi_R(\theta))\}_{\theta\in(0,\bar{\theta}]}$ is a valid graph satisfying the boundary conditions: i.e.,

(i) $\lim_{\theta\nearrow\infty}\big(q_L(\theta), \varphi_L(\theta)\big) = (0, 0)$.

(ii) $(q_R(\bar{\theta}), \varphi_R(\bar{\theta})) = (1, 0)$.

(iii) The left part and the right part meet at a point, i.e., $\lim_{\theta\searrow 0}\big(q_L(\theta), \varphi_L(\theta)\big) = \lim_{\theta\searrow 0}\big(q_R(\theta), \varphi_R(\theta)\big)$.

(iv) They have the same slope at the contact point, i.e., $\lim_{\theta\searrow 0}\frac{d\phi_L/d\theta}{dq_L/d\theta} = \lim_{\theta\searrow 0}\frac{d\phi_R/d\theta}{dq_R/d\theta}$.

First, observe that $\lim_{\theta\nearrow\infty} Z_L(\theta) = -\frac{2}{\pi}\lim_{\theta\nearrow\infty} K_{1/3}(\theta) = 0$ and $\lim_{\theta\nearrow\infty} Z_L'(\theta) = \frac{1}{\pi}\lim_{\theta\nearrow\infty}\big(K_{-2/3}(\theta) + K_{4/3}(\theta)\big) = 0$. Therefore, $\lim_{\theta\nearrow\infty} q_L(\theta) = 0$ and $\lim_{\theta\nearrow\infty}\varphi_L(\theta) = 0$, which proves (i).

Next, observe that the value of $\bar{\theta}$ was chosen to satisfy $Z_R(\bar{\theta}) = 0$, and the value of a was chosen to satisfy $q_R(\bar{\theta}) = a \times \bar{\theta}^{\frac{4}{3}}\left(Z_R'(\bar{\theta})\right)^2 = 1$, which proves (ii).

Finally, with some algebra, it can be shown that

$$\lim_{\theta\searrow 0} q_L(\theta) = \lim_{\theta\searrow 0} q_R(\theta) = \frac{2^{\frac{10}{3}}a}{3\Gamma^2(\frac{1}{3})}, \quad \lim_{\theta\searrow 0}\varphi_L(\theta) = \lim_{\theta\searrow 0}\varphi_R(\theta) = \frac{2^{\frac{8}{3}}b}{3\Gamma^2(\frac{2}{3})}, \tag{B.141}$$

and

$$\lim_{\theta \searrow 0} \frac{d\phi_L(\theta)/d\theta}{dq_L(\theta)/d\theta} = \lim_{\theta \searrow 0} \frac{d\phi_L(\theta)/d\theta}{dq_L(\theta)/d\theta} = \frac{3 \times 2^{\frac{1}{3}} b\Gamma(\frac{2}{3})}{a\Gamma(\frac{1}{3})}, \tag{B.142}$$

where Γ is the Gamma function. These results prove (iii) and (iv).

Some notes on the determination of constants. In the representation of the general solution, (B.139) and (B.140), there are seven constants to identify: a, b, $\bar{\theta}$, C_{L1}, C_{L2}, C_{R1}, and C_{R2}. The upper limit $\bar{\theta}$ and the constant a are uniquely determined by $(C_{L1}, C_{L2}, C_{R1}, C_{R2})$ due to (ii), and the constant b is also uniquely determined by a due to the identity $(b/a)^3 = 9/2$. We can also observe that the curve $\{(q(\theta), \varphi(\theta))\}_{\theta \in \mathbb{R}_+}$ is invariant to a uniform scaling of $(C_{L1}, C_{L2}, C_{R1}, C_{R2})$, and thus we can set $C_{R4} = -1$ without loss of generality. We can further obtain a system of equations from the other conditions: $C_{L2} = 0$ from (i), $C_{L1}^2\pi^2 = 4C_{R2}^2$ and $C_{L1}^2\pi^2 = 3C_{R1}^2 + 2\sqrt{3}C_{R1}C_{R2} + 3C_{R2}^2$ from (iii), and $C_{R2} = -\sqrt{3}C_{R1}$ from (iv). These equations uniquely determine the values of C_{L1}, C_{L2}, and C_{R1}. □